On the relationship between the Method of Least Squares and Gram-Schmidt orthogonalization

Hilmar Drygas, University of Kassel, Germany

Summary The method of Least Squares is due to Carl Friedrich Gauss. The Gram-Schmidt orthogonalization method is of much younger date. A method for solving Least Squares Problems is developed which automatically results in the appearance of the Gram-Schmidt orthogonalizers. Given these orthogonalizers an induction-proof is available for solving Least Squares Problems.

Keywords: Linear models, Method of Least Squares, simple regression, Steiner-theorem, orthogonal projection, Gram-Schmidt orthogonalization.

1 Introduction

The method of Least consister in the following problem. Given vectors $y, x_1, \ldots x_k \in \mathbb{R}^n$ find numbers $\beta_1, \ldots, \beta_k$ such that

$$\| y - \sum_{i=1}^{k} \beta_i x_i \| \tag{1}$$

is minimized. The underlying linear model is $y = x_1\beta_1 + \ldots + x_k\beta_k + \epsilon$, where ϵ is a disturbance term. Mostly, it is assumed that ϵ is a random vector with expectation 0 and covariance-matrix $\sigma^2 I_n$, where $\sigma > 0$ is unknown parameter. The method of Least Squares is therefore also desaribed by

$$\| \epsilon \| = \text{Min}\,. \tag{2}$$

The simplest linear model is $y = a\,1_n + \epsilon$, where 1_n is the all one-vector. The Least Squares Problem for estimating a can be solved by Steiners theorem.

1.1 Steiners Theorem:

$\sum\limits_{i=1}^{n} w_i(y_i - a)^2 = \sum\limits_{i=1}^{n} w_i(y_i - \overline{y}_{wgh})^2 + (\sum\limits_{i=1}^{n} w_i)(a - \overline{y}_{wgh})^2$, where $w_i \geq 0$, $\sum\limits_{i=1}^{n} w_i > 0$.

$y_{wgh} = (\sum\limits_{i=1}^{n} w_i)^{-1}(\sum\limits_{i=1}^{n} w_i y_i)$.

The proof follows from Pythagoras theorem since

$$\sum_{i=1}^{n} w_i(y_i - \overline{y}_{wgh})(a - \overline{y}_{wgh}) = 0\,. \tag{3}$$

Thus $a = \overline{y}_{wgh}$ solves the Least Squares Problem $\sum_{i=1}^{n} w_i(y_i - a)^2 = \mathrm{Min}\,.$

This theorem can also be used to solve the regression model $y_i = \alpha + \beta x_i + \epsilon_i\,,\ i = 1, \ldots, n\,,\ y = \alpha 1_n + \beta x + \epsilon\,.$ The task consists in minimizing

$$Q = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2\,. \tag{4}$$

By Steiners Theorem we get the solution

$$\hat{\alpha} = \overline{y} - \beta\overline{x}\,,\ \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,. \tag{5}$$

By plugging in we get

$$\begin{aligned} Q &= \sum_{i=1}^{n}\Big(y_i - \overline{y} - \beta(x_i - \overline{x})\Big)^2 \\ &= \sum_{i=x_i\neq\overline{x}} (x_i - \overline{x})^2 \left(\frac{y_i - \overline{y}}{x_i - \overline{x}} - \beta\right)^2 + \sum_{i:x_i=\overline{x}} (y_i - \overline{y})^2\,. \end{aligned} \tag{6}$$

According to Steiners theorem the minimizing β is a weighted mean of the slopes $\frac{y_i - \overline{y}}{x_i - \overline{x}}$, namely

$$\hat{\beta} = \frac{\sum\limits_{i=x_i\neq\overline{x}} (x_i - \overline{x})^2 \frac{(y_i - \overline{y})}{(x_i - \overline{x})}}{\sum\limits_{i=x_i\neq\overline{x}} (x_i - \overline{x})^2} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}\,. \tag{7}$$

If all x_i are equal to $\overline{x}$, then β is arbitrary since Q does not depend on β.

This method of successive solution and plugging in can be extended to the general case as will be shown in the next section.

2 Generalization of successive estimation

2.1 Generalized Steiner Theorem:

$\| y - ax \|^2 = \| y - \frac{(y,x)}{(x,x)} x \|^2 + \| x \|^2 \left(a - \frac{(x,y)}{(x,x)} \right)^2$ if $x \neq 0$.

Proof: $\left(y - \frac{(y,x)}{(x,x)} x \right)$ and x are orthogonal. Pythagoras Theorem therefore yields the result. □

2.2 Corollary:

$a = \frac{(y,x)}{(x,x)}$ is the Least Squares-solution of $\| y - ax \| =$ Min. and $a = 0$ yields a proof of the Cauchy-Schwarz inequality.

Now we want to minimize

$$\| y - \sum_{i=1}^{k} \beta_i x_i \|^2 \quad . \tag{1}$$

If $x_1 = 0$, then β_1 does not appear in (2.1) and it is therefore arbitrary. If $x_1 \neq 0$, then according to theorem 2.1

$$\hat{\beta}_1 = \frac{(y - \sum\limits_{i=2}^{k} \beta_i x_i, x_1)}{(x_1, x_1)} \, . \tag{2}$$

By plugging in we get the new minimization problem

$$\| y^{(2)} - \sum_{i=2}^{k} \beta_i x_i^{(2)} \| = \text{ Min} \, , \tag{3}$$

where

$$y^{(2)} = y - \frac{(y, x_1)}{(x_1, x_1)} x_1 = P_{\{x_1\}^\perp} y, x_i^{(2)} = x_i - \frac{(x_i, x_1)}{(x_1, x_1)} x_1 = P_{\{x_1\}^\perp} x_i \, . \tag{4}$$

If $x_2^{(2)} \neq 0$ - otherwise β_2 is arbitrary - we obtain

$$\hat{\beta}_2 = \frac{(y^{(2)} - \sum\limits_{i=3}^{k} \beta_i x_i^{(2)}, x_2^{(2)})}{(x_2^{(2)}, x_2^{(2)})} = \frac{(y - \sum\limits_{i=3}^{k} \beta_i x_i, x_2^{(2)})}{(x_2^{(2)}, x_2^{(2)})} \, . \tag{5}$$

and again by plugging in we get a new problem with $y^{(3)}, x_i^{(3)}, i = 3, \ldots, k$. Continuing we get successively the solutions $(j = 3, \ldots, k)$

$$\hat{\beta}_j = \frac{(y - \sum\limits_{i=j+1}^{k} \beta_i x_i^{(j)}, x_j^{(j)})}{(x_j^{(j)}, x_j^{(j)})} \, , \text{ if } x_j^{(j)} \neq 0 \tag{6}$$

and finally

$$\hat{\beta}_k = \frac{(y, x_k^{(k)})}{(x_k^{(k)}, x_k^{(k)})} \, , \text{ if } x_k^{(k)} \neq 0 \, . \tag{7}$$

In order to simplity the notation we define

$$q_1 = x_1 \, , \, q_j = x_j^{(j)} \, , \, j = 2, \ldots, k \, . \tag{8}$$

Then

$$\begin{aligned} x_i^{(l)} &= x_i^{(l-1)} - \frac{(x_i^{(l-1)}, q_{l-1})}{(q_{l-1}, q_{l-1})} q_{i-1} \\ &= P_{\{q_{l-1}\}^\perp} x_i^{(l-1)} \, , \, i = l, \ldots, k \, , l = 1, \ldots, k_i \end{aligned} \tag{9}$$

where, of course, $x_i^{(1)} = x_i \, , \, i = 1, \ldots, k$. Therefore

$$q_l = P_{\{q_{i-1}\}^\perp} x_i^{(l-1)} \tag{10}$$

and

$$x_i^{(l)} = P_{\{q_{l-1}\}^\perp} P_{\{q_{l-2}\}^\perp} \ldots P_{\{q_1\}^\perp} x_i \tag{11}$$

$$q_l = P_{\{q_{l-1}\}^\perp} \ldots P_{\{q_l\}^\perp} x_l \, , \, l = 2, \ldots, k \, . \tag{12}$$

The next step consists in proofing that

$$\prod_{j=1}^{i-1} P_{\{q_{i-j}\}^\perp} = P_{\{q_1, \ldots, q_{i-1}\}^\perp} \, . \tag{13}$$

By Achieser/Glasmann, 1981, p. 97 pp. the product of projections is a projector iff the projectors commute. By page 189 in Rao/Mitra, 1971, the projection onto the intersection of the subspaces M and N is given by

$$2P(P+Q)^- Q \tag{14}$$

where P is the projection onto M and Q the projection onto N. There must be a simple formula for the generalized inverse of $(P+Q)$, namely $(P+Q)^+$. This formula will be given by the following theorem:

2.3 Theorem:

If $PQ = QP$, then $(P+Q)^+ = P + Q - \frac{3}{2} PQ$.

Proof: The proof follows from verification. An alternative is that P and Q are jointly diagonalizable if $PQ = QP$. $P = C \operatorname{diag}(\lambda_1, \ldots, \lambda_n) C'$, $Q = C \operatorname{diag}(\mu_1, \ldots, \mu_n) C'$ and the λ_i and μ_i are either 0 or 1. Then $P + Q = C\,(\operatorname{diag}(\lambda_1 + \mu_i), \ldots, (\lambda_\mu + \mu_n)) C$,
$(P+Q)^+ = C \operatorname{diag}((\lambda_1 + \mu_1)^+, \ldots, (\lambda_n + \mu_n)^+) C'$. But

$$(\lambda_i + \mu_i)^+ = \lambda_i + \mu_i - \frac{3}{2} \lambda_i \mu_i \tag{15}$$

in all possible cases. $\square$

2.4 Theorem:

PQ is the projection onto im $(P)\cap$ im (Q) iff $QM^\perp \subseteq M^\perp$. Sufficient for this is $M^\perp \subseteq N$.

Proof: PQ ist the projection onto $M \cap N$ iff it is the identity on $N \cap M$ and vanishes on $(M \cap N)^\perp = M^\perp + N^\perp$ Since the other properties are obvious only $PQM^\perp = 0$ must be considered. This is equivalent to $QM^\perp \subset M^\perp$. This condition met if $M^\perp \subseteq N$. □

2.5 Theorem:

$$\prod_{j=1}^{i-1} P_{\{q_{i-j}\}^\perp} = P_{\{q_1,\ldots,q_{i-1}\}^\perp} \quad \text{and} \quad q_i \in \{q_1, \ldots, q_{i-1}\}^\perp . \tag{16}$$

Proof: Mathematical induction. The first assertion of the theorem is correct for $i = 2$ and $q_2 = P_{\{q_1\}^\perp} x_2 \in \{q_1\}^\perp$. Let by induction assumption

$$\prod_{j=1}^{i-1} P_{\{q_j\}^\perp} = P_{\{q_1,\ldots,q_{i-1}\}^\perp} \quad \text{and} \quad q_i \in \{q_1, \ldots, q_{i-1}\}^\perp . \tag{17}$$

Then

$$\prod_{j=1}^{i} P_{\{q_{i-j}\}^\perp} = P_{\{q_i\}^\perp} P_{\{q_1,\ldots,q_{i-1}\}^\perp} . \tag{18}$$

Since $q_i \in \{q_1, \ldots, q_{i-1}\}^\perp$ it follows from theorem $(M = \{q_i\}^\perp, M^\perp = \{\lambda_{q_i}; \lambda \in \mathbb{R}\})$ that

$$\prod_{j=1}^{i} P_{\{q_j\}^\perp} = P_{\{q_i\}^\perp} P_{\{q_1,\ldots,q_{i-1}\}^\perp} = P_{\{q_i\}^\perp \cap \{q_1,\ldots,q_{i-1}\}^\perp} = P_{\{q_1,\ldots,q_i\}^\perp} . \tag{19}$$

Since $q_{i+1} = P_{\{q_1,\ldots,q_i\}^\perp} x_{i+1} \in \{q_1, \ldots, q_i\}^\perp$ also the second assertion is proved. □

2.6 Corollary:

If $q_0 = 0$, then $q_i = P_{\{q_0,\ldots,q_{i-1}\}^\perp} x_{i,i=1,\ldots,k}$ and $x_i^{(l)} = P_{\{q_0,q_1,\ldots,q_{l-1}\}^\perp} x_i$.

Since

$$q_i = P_{\{q_0,\ldots,q_{i-1}\}^\perp} x_i = x_i - P_{\text{span}\{q_1,\ldots,q_{i-1}\}} x_i = x_i - \sum_{j:q_j \neq 0}^{i-1} \frac{(q_j, x_i)}{(q_j, q_j)} q_j \tag{20}$$

the q_i desoribe the Gram-Schmidt orthogonalization procedure. It follows that from the principle of Least Squares the Gram-Schmidt orthogonalization procedure could be invented.

3 An induction proof

Since the Gram-Schmidt orthogonalization procedure is well-known now the Least Squares Solutions can also be proved by mathematical induction. The induction ist on the number m of linear independent vectors among $x_1, \ldots, x_k$. We assume that $x_1, \ldots, x_m$ are linearly independent and Rank $\{x_1, \ldots, x_k\} = m$. There fore $x_{m+1}, \ldots, x_k$ are linear combinations of $x_1, \ldots, x_m$. As we habe seen in the last section

$$\hat{\beta}_1 = \frac{(y - \sum\limits_{i=2}^{k} \beta_i x_i, x_1)}{(x_1, x_1)} \tag{1}$$

and by plugging in we get the new minimization problem

$$\text{Minimize} \quad \| \, y^{(2)} - \sum_{i=2}^{k} \beta_i x_i^{(2)} \, \|^2 \tag{2}$$

where

$$y^{(2)} = P^{\perp}_{\{x_1\}} y \, , \; x_i^{(2)} = P^{\perp}_{\{x_1\}} x_i \, , \; i = 2, \ldots, k \, . \tag{3}$$

3.1 Lemma:

Let Rank $(x_1, \ldots, x_k)$ be equal m and let $x_1, \ldots, x_m$ be linearly independent. Then $x_i^{(2)}$, $i = 2, \ldots, m$ are linearly independent and the $x_i^{(2)}$, $i > m$ are linear combinations of the $x_i^{(2)}$, $i = 2, \ldots, m$.

Proof:

a) From $\sum\limits_{i=2}^{m} \lambda_i x_i^{(2)} = P(\sum\limits_{i=2}^{m} \lambda_i x_i) = 0$ – we write P for short instead of $P_{\{x_1\}^\perp}$ – it follows that $\sum\limits_{i=2}^{m} \lambda_i x_i \in \text{span}\{x_1\}$ and hence $\lambda_2 = \ldots = \lambda_m = 0$ from the linear independence of $x_1, \ldots, x_m$.

b) For $i > m$ we get $x_i^{(2)} = \sum\limits_{j=2}^{m} \lambda_{ij} x_j^{(2)}$ if $x_i = \sum\limits_{j=1}^{m} \lambda_{ij} x_j$. $\square$

3.2 Theorem:

Let Rank $(x_1, \ldots, x_k) = m$ and let, moreover, $x_1, \ldots, x_m$ be linearly independent. Futhermore, let $q_1, \ldots, q_m$ be the pairwise orthogonal vectors obtained from $(x_1, \ldots, x_m)$ by applying the Gram-Schmidt orthogonalization procedure. Then the Least Squares solutions $\hat{\beta}_1, \ldots, \hat{\beta}_m$ are recursively given by

$$\hat{\beta}_m \;=\; \frac{(q_m, y - \sum\limits_{i=m+1}^{k} \beta_i x_i)}{(q_m, q_m)} \tag{4}$$

$$\hat{\beta}_i \quad = \quad \frac{(q_i, y - \sum\limits_{j=i+1}^{m} \hat{\beta}_j x_j - \sum\limits_{j=m+1}^{k} \beta_j x_j)}{(q_i, q_i)} \tag{5}$$

$i = m-1\,,\ m-2, \ldots, 1\,.$ Here $\beta_{m+1}, \ldots, \beta_k$ are completely arbitrary.

Moreover

$$y - \sum_{i=1}^{m} \hat{\beta}_i x_i - \sum_{j=m+1}^{k} \beta_i x_i$$

does not depend on $\beta_{m+1}, \ldots, \beta_k\,.$

Proof: Mathematical induction on $m\,.$ If $m = 1\,,$ then

$$\hat{\beta}_1 = \frac{(x_1, y - \sum\limits_{i=2}^{k} \beta_i x_i)}{(x_1, x_1)} \tag{6}$$

and

$$y - \hat{\beta}_1 x_1 - \sum_{i=2}^{k} \beta_i x_i = y^{(2)} - \sum_{i=2}^{k} \beta_i x_i^{(2)}\,. \tag{7}$$

But since $x_i \in \operatorname{span}\{x_1\}$ it follows that $x_i^{(2)} = 0\,,\ i = 2, \ldots, k$ and therefore

$$y - \hat{\beta}_1 x_1 - \sum_{i=2}^{k} \beta_i x_i = y^{(2)} \tag{8}$$

which does not depend on $\beta_2, \ldots, \beta_k\,.$

We now arrive at the problem of minimizing

$$\parallel y^{(2)} - \sum_{i=2}^{k} \beta_i x_i^{(2)} \parallel\ . \tag{9}$$

By the induction assumption using that $x_2^{(2)}, \ldots, x_2^{(m)}$ are linearly independent and Rank $(x_2^{(2)}, \ldots, x_k^{(2)}) = m-1$ we get that the solutions are as follows:

$$(q_m^{(2)}, q_m^{(2)})\, \hat{\beta}_m = (q_m^{(2)}, y^{(2)} - \sum_{i=m+1}^{k} \beta_i x_i^{(2)}) \tag{10}$$

and

$$(q_i^{(2)}, q_i^{(2)})\, \hat{\beta}_i = (q_i^{(2)}, y^{(2)} - \sum_{j=i+1}^{m} \hat{\beta}_j x_j - \sum_{j=m+1}^{k} \beta_j x_j) \tag{11}$$

$i = m-1, \ldots, 2$ Here $\beta_{m+1}, \ldots, \beta_k$ are arbitrary numbers and $q_2^{(2)}, \ldots, q_m^{(2)}$ are obtained by applying the Gram-Schmidt orthogonalization procedure to $x_2^{(2)}, \ldots, x_m^{(2)}\,.$ Moreover, $y^{(2)} - \sum\limits_{i=2}^{m} \hat{\beta}_i \hat{x}_i - \sum\limits_{i=m+1}^{k} \beta_i x_i$ does not depend on $\beta_{m+1}, \ldots, \beta_k\,.$ From this it follows that

$$y - \sum_{i=1}^{m} \hat{\beta}_i x_i - \sum_{j=m+1}^{k} \beta_i x_i = y^{(2)} - \sum_{i=2}^{k} \hat{\beta}_i x_i^{(2)} - \sum_{i=m+1}^{k} \beta_i x_i^{(2)} \tag{12}$$

as well does not depend on $\beta_{m+1}, \ldots, \beta_k$.

We now prove by mathematical induction that $q_i^{(2)} = q_i$, $i = 2, \ldots, m$. This is indeed correct for $i = 2$ since $x_2^{(2)} = q_2^{(2)} = x_2 - \frac{(x_1, x_2)}{(x_1, x_1)} x_1 = q_2$ and by using the induction assumption we get

$$q_i^{(2)} = x_2^{(2)} - \sum_{j=2}^{i-1} \frac{(x_i^{(2)}, q_j^{(2)})}{(q_j^{(2)}, q_j^{(2)})} q_j^{(2)} = x_i - \frac{(x_i, x_1)}{(x_1, x_1)} x_1 - \sum_{j=2}^{i-1} \frac{(x_i^{(2)}, q_j)}{(q_j, q_j)} q_j \,. \tag{13}$$

Since $(x_i^{(2)}, q_j) = (x_i, q_j)$ for $j \geq 2$, it follows indeed that $q_i^{(2)} = q_i$, $i = 2, \ldots, m$.

From $(q_m, q_1) = 0$ for $i \geq 2$ we finally get

$$\hat{\beta}_m = \frac{(q_m, y^{(2)} - \sum\limits_{i=m+1}^{k} \beta_i x_i^{(2)})}{(q_m, q_m)} = \frac{(q_m, y - \sum\limits_{i=m+1}^{k} \beta_i x_i)}{(q_m, q_m)} \tag{14}$$

and for $i = m-1, \ldots, 2$

$$\begin{aligned} \hat{\beta}_i &= \frac{(q_i, y^{(2)} - \sum\limits_{j=i+1}^{m} \hat{\beta}_j x_j^{(2)} - \sum\limits_{j=m+1}^{k} \beta_j x_j^{(2)})}{(q_i, q_i)} \\ &= \frac{(q_i, y - \sum\limits_{j=i+1}^{m} \hat{\beta}_j x_j - \sum\limits_{j=m+1}^{k} \beta_j x_j)}{(q_i, q_i)} \,. \end{aligned} \tag{15}$$

This is completed by

$$\hat{\beta}_1 = \frac{(x_1, y - \sum\limits_{j=2}^{m} \hat{\beta}_j x_j - \sum\limits_{j=m+1}^{k} \beta_j x_j)}{(x_1, x_1)} = \frac{(q_1, y - \sum\limits_{j=2}^{m} \hat{\beta}_j x_j - \sum\limits_{j=m+1}^{k} \beta_j x_j)}{(q_1, q_1)} \,. \tag{16}$$

$\square$

3.3 References

Achieser-Glasmann (1981), Theorie der linearen Operatoren im Hilbert-Raum, Akademie-Verlag Berlin.

Drygas, H. (2008), QR-decomposition from the statistical point of view, Recent Advances in Linear Models and Related Areas, Essays in Honour of Helge Toutenburg, Shalab and Heumann (Eds) p. 293-311, Physica-Verlag, Springer, Heidelberg.

Drygas, H. (2009), Statistical Analysis of Diabetes Mellitus. Discussiones Mathematicae, Probability and Statistics, 29 (2009), p. 69 - 90, University of Zielona Góra, Poland.

Gauss, C. F. "Theoria combinationis Observationum Erroribus Minimis Obnoxae" Comm. recent. Soc. Gött. 5 (1819 - 1822) oder Werke, Bd IV, Leipzig, 1879, p. 3 - 53. A German translation under the title "Abhandlungen zur Methode der kleinsten Quadrate" was edited by Börsch and Simon in 1987, Druck and Verlag von P. Stankiewicz Buchdruckerei, Berlin.

Linnik, J. W. (1961), Methode der kleinsten Quadrate in Moderner Darstellung, VEB Deutscher Verlag der Wissenschaften, Berlin.

Rao, C. R. and Mitra, S. K. (1971), Generalized Inverse of Matrices and its Applications, Wiley, New York - London - Sydney - Toronto.

Richter, H., Mammitzsch, V. (1973), Methode der kleinsten Quadrate mit Übungen und Aufgaben.