Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **50**, 2 (2011) 103–110

Strange Design Points in Linear Regression[*]

Andrej PÁZMAN

Department of Applied Mathematics and Statistics
Faculty of Mathematics, Physics and Informatics
Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia
e-mail: pazman@fmph.uniba.sk

Dedicated to Lubomír Kubáček on the occasion of his 80th birthday

(Received March 31, 2011)

Abstract

We discuss, partly on examples, several intuitively unexpected results in a standard linear regression model. We demonstrate that direct observations of the regression curve at a given point can not be substituted by observations at two very close neighboring points. On the opposite, we show that observations at two distant design points improve the variance of the estimator. In an experiment with correlated observations we show somewhat unexpected conditions under which a design point gives no or very little information about the estimated parameters, and so can be excluded from the design. For completeness we repeat briefly known conditions under which a design point is sensitive to the presence of outliers.

Key words: singular models, optimal design, correlated observations

2010 Mathematics Subject Classification: 62K05, 62J05

1 Introduction

We consider a standard linear regression model

$$y_x = f^T(x)\,\theta + \varepsilon_x, \quad x \in \mathcal{X},\ \theta \in \mathbb{R}^p \tag{1}$$

with data $y_{x_1}, \ldots, y_{x_N}$ observed in N design points $x_1, \ldots, x_N$ taken from a design space $\mathcal{X}$. In a vector notation we have $y = F\theta + \varepsilon$ with $y = (y_{x_1}, \ldots, y_{x_N})^T$,

[*] Supported by the VEGA grant No. 1/0077/09.

$F_{ij} = f_j(x_i) =$ elements of a known design matrix. The vector of random errors $\varepsilon = (\varepsilon_{x_1}, \ldots, \varepsilon_{x_N})^T$ is supposed to have $E(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 W$ with W a known positive definite matrix.

A well known example is a *polynomial regression* on an interval. Then $\mathcal{X} = [a, b]$,

$$y_x = \theta_1 + \theta_2 x + \cdots + \theta_m x^{m-1} + \varepsilon_x$$

hence $f^T(x) = \left(1, x, \ldots, x^{m-1}\right)$.

The best linear unbiased estimator (BLUE) of the vector θ is equal to

$$\hat{\theta} = M^{-1} F^T W^{-1} y \tag{2}$$

Here

$$M \equiv M(x_1, \ldots, x_N) = F^T W^{-1} F$$

is the information matrix (for $\sigma = 1$). The variance matrix of $\hat{\theta}$ is $\mathrm{Var}(\hat{\theta}) = \sigma^2 M^{-1}$ if M is nonsingular, and $\hat{\theta}$ is not unique if M is singular, but still we have in this case a BLUE for $h^T\theta$, with h a given vector such that $h \in \mathcal{M}(M) =$ the column space of the matrix M. Then

$$\mathrm{Var}\left(h^T \hat{\theta}\right) = \sigma^2 h^T M^- h$$

where M^- is an arbitrary g-inverse of M, and $\hat{\theta} = M^- F^T W^{-1} y$ is any solution of the normal equation $M\theta = F^T W^{-1} y$.

Evidently, the position of the design points $x_1, \ldots, x_N$ influences the variances, as well as the form of the estimated regression function

$$\eta\left(x, \hat{\theta}\right) \equiv f^T(x)\hat{\theta}$$

Still, it seems that nothing surprising can be found in this model. We shall try to prove the opposite.

2 Design points which cannot be approximated by neighboring design points

Take $[a, b] = [0, 10]$ and

$$\eta(x, \theta) = \theta_1 x + \theta_2 x^2 \tag{3}$$

a quadratic polynomial without intercept. Suppose that we have to perform 10 independent observation with the same unknown variance σ^2, and the aim is to estimate the value of this polynomial at the point $\bar{x} = 1$.

Observations at one point: Apparently, a “natural” design of the experiment would be to perform all 10 observations at the same design point $\bar{x} = 1$, i.e. in our previous notation, $x_1 = \cdots = x_{10} = \bar{x}$. We have then evidently

$$\mathrm{Var}\left(\eta\left(\bar{x}, \hat{\theta}\right)\right) = \mathrm{Var}\left(\frac{1}{10}\sum_{i=1}^{10} y_i\right) = \frac{\sigma^2}{10}$$

Observations at two neighboring points: Suppose that for some reason the experimenter wants to divide the observations into two groups: Five observations at the point $x_1 = \bar{x} + t$ and five at $x_2 = \bar{x} + ct$, with $t > 0$ but small, and with $c \in [-1, 1]$. Then direct computation gives

$$M^{-1}(x_1, x_2) = \frac{1}{5x_1^2 x_2^2 (x_1 - x_2)^2} \begin{pmatrix} x_1^4 + x_2^4 & -x_1^3 - x_2^3 \\ -x_1^3 - x_2^3 & x_1^2 + x_2^2 \end{pmatrix}$$

and after some standard computation, since

$$\eta\left(\bar{x}, \hat{\theta}\right) = \hat{\theta}_1 + \hat{\theta}_2,$$

the variance of the BLUE for $\eta(\bar{x}, \theta)$ is equal to

$$\mathrm{Var}\Big(\eta\Big(\bar{x}, \hat{\theta}\Big)\Big) = \sigma^2 \begin{pmatrix} 1 & 1 \end{pmatrix} M^{-1}(x_1, x_2) \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{\sigma^2}{5(1-c)^2} \left[\frac{1}{(1+ct)^2} + \frac{c^2}{(1+t)^2}\right]$$

which for $t \to 0$ tends to

$$\frac{\sigma^2 (1 + c^2)}{5(1-c)^2}$$

Consequently, $\mathrm{Var}\left(\eta\left(\bar{x}, \hat{\theta}\right)\right)$ *is arbitrarily large if we take c sufficiently close to* 1*, regardless of how small is t, i.e. how close are the points* x_1*, and* x_2 *to* $\bar{x} = 1$. And for any $c \neq -1$ the limit variance is larger than when performing all observations at one point $\bar{x}$.

Consequences:

i) *Mathematically*, we have a very clear discontinuity of the variance when considering it as a function of the design. This discontinuity is discussed on a theoretical level in Pázman (1986), p. 63–69.

ii) *We have a conflict between singular and regular regression models:* If we observe just at one point $\bar{x}$, we have a singular regression model, which can be used, since we suppose a priori that just one parameter is important, namely $\eta(\bar{x}, \theta)$. On the other hand, if we observe at two points, we have a regular model with two parameters to be estimated. But if the points x_1, x_2 are very close to $\bar{x}$, the model is "bad conditioned", nearly singular, and some functions of the parameters are estimated with high variances. Accidently, it is the parameter $\eta(\bar{x}, \theta)$ in our case.

iii) *We have a statistical paradox on an elementary level:* In fact, if x_1, x_2 are very close to $\bar{x}$, in practice one can not be sure whether we observed at $\bar{x}$ or at x_1, and x_2. So the estimator $\eta\big(\bar{x}, \hat{\theta}\big)$ should not be so sensitive to the choice of design points as given by the theory. However, this reflection contains implicitly the *a priori assumption* that the values $\eta(\bar{x}, \theta)$, $\eta(x_1, \theta)$, $\eta(x_2, \theta)$ do not differ very much. But our quadratic model (3) is not adequate for such an assumption, since it does not exclude a very narrow parabola, giving very different values of $\eta(\bar{x}, \theta)$, $\eta(x_1, \theta)$, $\eta(x_2, \theta)$, even if $\bar{x}$, x_1, x_2 are very close to each other. *So the "paradoxical" result obtained from the theory is statistically correct, but the model does not correspond to our "intuitive" assumptions.*

To solve the problem, either we must use a Bayesian modelling rejecting "a priori" the possibility of a narrow parabola, or we must use another regression model, say we must suppose that for x in a neighborhood of the point $\bar{x}$ a one-parameter model

$$\eta(x, \theta) = \theta_1$$

is acceptable, or simply we must accept that approximations by singular models are sometimes preferable. This may be an argument to use singular models largely considered in the papers of L. Kubáček (cf. e.g. Fišerová et al (2007)).

The situation is even worse when we want to estimate a nonlinear function of θ_1 and θ_2. For example, suppose that in the model (3) we want to estimate the position x_o of the extreme point of the parabola $\theta_1 x + \theta_2 x^2$. By taking

$$d\left[\theta_1 x + \theta_2 x^2\right]/dx = 0$$

we obtain $x_o = -\frac{\theta_1}{2\theta_2}$, so x_o should be estimated by $-\frac{\hat{\theta}_1}{2\hat{\theta}_2}$. As is proven in Pázman and Pronzato (2006), if we observe m times at the "true" point x_o and $N - m$ times at a point x_1 which is close to x_o, it may be that the limit distribution of $-\frac{\hat{\theta}_1}{2\hat{\theta}_2}$ is asymptotically not normal for $N \to \infty$, or still normal but with a "strange" variance or with a "strange" speed of convergence. All depends on the behavior of m/N.

So it is better to do all 10 observations at the point $\bar{x} = 1$ than to split them into two neighbor design points. But what about to split the observations into two distant points?

Take two design points: $x_1 = 4.14$, and $x_2 = 10$, and take 8 observations at x_1 and 2 observations at x_2. Straightforward calculus gives

$$M = M(x_1, x_2) = \begin{pmatrix} 337.12 & 2567.7 \\ 2567.7 & 22350.1 \end{pmatrix}$$

and

$$\mathrm{Var}\left[\eta\left(\bar{x}, \hat{\theta}\right)\right] = \sigma^2 \left(\, 1 \;\; 1 \,\right) M^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = (0.186) \times \frac{\sigma^2}{10}$$

which is much less then when we perform all 10 observations at $\bar{x}$. Notice that we can obtain the two design points x_1, x_2 when using for our case the graphical method for computing optimum designs based on the Elfving's theorem, cf. e.g. Pázman (1986), p. 71.

3 Design points giving zero or little information about θ

Consider the design $x_1, \ldots, x_N$ in the general model (1). Suppose first that the observations are uncorrelated, with $W = I$. Then the information matrix is

$$M(x_1, \ldots, x_N) = \sum_{i=1}^{N} f(x_i) f^T(x_i)$$

So, the information contained in the observation at one design point x_i is equal to $f(x_i)f^T(x_i)$. As a consequence, *in the uncorrelated case a design point x_i gives zero information about θ if and only if $f(x_i) = 0$.*

The situation may be quite different, and somehow surprising, when the observations are correlated, $\mathrm{Var}(\varepsilon) = \sigma^2 W$, with $W \neq I$. Now

$$M(x_1, \dots x_N) = F^T W^{-1} F = \sum_{i=1}^{N} \sum_{j=1}^{N} f(x_i) \left\{W^{-1}\right\}_{ij} f^T(x_j) \qquad (4)$$

Intuitively, one can perhaps argue also here that $f(x) = 0$ implies that in the model

$$y_x = f^T(x)\,\theta + \varepsilon_x, \quad x \in \mathcal{X}$$

y_x is not influenced by the value of θ, hence the observation at x should give no information about θ. This intuitive approach is false. To see this, consider an example.

Example 1 Suppose that $\theta \in \mathbb{R}$, take $\{x_1, x_2\}$ a two point design such that $f(x_1) = 0$, $f(x_2) = 1$, and suppose that $W_{1,1} = W_{2,2} = 1$, but $W_{1,2} \neq 0$. Then, according to (4)

$$M(x_1, x_2) = (0,1) \begin{pmatrix} 1 & W_{1,2} \\ W_{1,2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{1 - (W_{1,2})^2} > 1 = M(\{x_2\}).$$

So although $f(x_1) = 0$, by deleting the point x_1 from the design we can lose much information. The contribution of the point x_1 to $M(x_1, x_2)$ is very large, even if $f(x_1) = 0$, when the observations y_{x_1} and y_{x_2} are highly correlated.

To express that the information at a design point x_k is small we need measures of information which are one dimensional (scalars). Like in experiments with uncorrelated observations, we shall consider information functionals, which are concave, monotone, real-valued functions defined on the set of positive definite matrices, cf. Pukelsheim (1993) for justification of these properties, or Pázman (1986) for a statistical interpretation. The gradient of such a functional Φ is the matrix $\nabla\Phi[M]$ of the same dimension as M, with components

$$\{\nabla\Phi[M]\}_{ij} = \frac{\partial\Phi[M]}{\partial\{M\}_{ij}}.$$

Well known examples are the D-optimality functional $\Phi[M] = \ln\det(M)$ with $\nabla \ln\det(M) = M^{-1}$, or the A-optimality functional $\Phi[M] = -\operatorname{tr}\left(M^{-1}\right)$ with $\nabla\left[-\operatorname{tr}\left(M^{-1}\right)\right] = M^{-2}$.

For a fixed design $D = \{x_1, \dots, x_N\}$ we denote

$$a(x_k) = \sum_{i=1}^{N} \left\{W^{-1}\right\}_{x_k, x_i} f(x_i)$$

and

$$\|a(x_k)\|_\Phi^2 = a^T(x_k)\,\nabla\Phi[M(D)]\,a(x_k),$$

which is a (pseudo)norm, since concavity of Φ implies that the gradient $\nabla\Phi\left[M\right]$ is a positive (semi)definite matrix of M. For example, if $\Phi\left(M\right) = \ln\det(M)$, then $\left\|a\left(x_k\right)\right\|_\Phi^2 = a^T\left(x_k\right)\left[M\left(D\right)\right]^{-1} a\left(x_k\right)$.

Proposition 1 *Suppose that the point x_k is deleted from the design $D = \{x_1, \ldots, x_N\}$. Then the resulting information matrix is*

$$M\left(D - \{x_k\}\right) = M\left(D\right) - \frac{a\left(x_k\right) a^T\left(x_k\right)}{\{W^{-1}\}_{x_k,x_k}}$$

and

$$\Phi\left[M\left(D - \{x_k\}\right)\right] = \Phi\left[M\left(D\right)\right] - \frac{\left\|a\left(x_k\right)\right\|_\Phi^2}{\{W^{-1}\}_{x_k,x_k}} + o\left(\left\|a\left(x_k\right)\right\|_\Phi^3\right)$$

where

$$\lim_{t\to 0} o\left(t\right)/t = 0.$$

Proof The expression for $M\left(D - \{x_k\}\right)$ has been derived in Pázman (2010). Further, from the Taylor formula applied to the matrix function $M \to \Phi(M)$ we obtain

$$\begin{aligned}
&\Phi\left[M(D) - \frac{a\left(x_k\right) a^T\left(x_k\right)}{\{W^{-1}\}_{x_k,x_k}}\right] = \\
&= \Phi\left[M\left(D\right)\right] - \mathrm{tr}\left\{\nabla\Phi\left[M\right]_{M(D)}\left[\frac{a\left(x_k\right) a^T\left(x_k\right)}{\{W^{-1}\}_{x_k,x_k}}\right]\right\} + o\left(\left\|a\left(x_k\right)\right\|_\Phi^3\right) \\
&= \Phi\left[M\left(D\right)\right] - \frac{\left\|a\left(x_k\right)\right\|_\Phi^2}{\{W^{-1}\}_{x_k,x_k}} + o\left(\left\|a\left(x_k\right)\right\|_\Phi^3\right)
\end{aligned}$$

□

Consequently, the amount of information about θ obtained from a design point x_k is small iff the expression $\frac{\|a(x_k)\|_\Phi^2}{\{W^{-1}\}_{x_k,x_k}}$ is small, and it is zero if and only if $a\left(x_k\right) = 0$, as follows from the first equality in Proposition 1. Indeed, a symmetric matrix A is zero if and only if $u^T A u = 0$ for every vector u. So, design points x_k giving a small value of $\frac{\|a(x_k)\|_\Phi^2}{\{W^{-1}\}_{x_k,x_k}}$ can be excluded from the design D without essential loss of information about the parameters of the regression model.

Notice, that the extreme case of zero information at x_k has been considered yet in Näther (1985) where instead of $a\left(x_k\right) = 0$ another but equivalent condition is used.

4 Design points sensitive to outliers

For completeness, we present here a known result from the diagnostics of a linear model (cf. Zvára (1989)).

We have the following statement.

Proposition 2 *Suppose that the information matrix M is nonsingular. Then for any $i = 1, \ldots, N$ we have*

$$f^T(x_i)\, M^{-1} f(x_i) \leq \{W\}_{ii}$$

In the extreme case that

$$f^T(x_i)\, M^{-1} f(x_i) = \{W\}_{ii}$$

the graph of the estimated regression function $x \in [a, b] \to \eta\left(x, \hat{\theta}\right)$ contains the point (x_i, y_{x_i}).

Simple proof According to (2) we have

$$\mathrm{Var}\left[y - F\hat{\theta}\right] = \mathrm{Var}\left[\left(I - FM^{-1}F^T W^{-1}\right) y\right] = \sigma^2(W - FM^{-1}F^T)$$

Consequently

$$f^T(x_i)\, M^{-1} f(x_i) = \left\{FM^{-1}F^T\right\}_{ii} \leq \{W\}_{ii}\,.$$

If $\left\{FM^{-1}F^T\right\}_{ii} = \{W\}_{ii}$, then

$$E\left[\left(y_i - F_{i.}\hat{\theta}\right)^2\right] = \mathrm{Var}\left[y_i - F_{i.}\hat{\theta}\right] = 0,$$

hence

$$y_i = F_{i.}\hat{\theta} = \eta\left(x_i, \hat{\theta}\right)$$

with probability one. □

Corollary 1 *If $f^T(x_i)\, M^{-1} f(x_i)$ is close to $\{W\}_{ii}$, then, even before performing the experiment, we know that the whole estimated regression function is strongly influenced by y_{x_i}, even if y_{x_i} is an outlier.*

Remark 1 In the case of uncorrelated observations with constant variances this emphasizes the importance of using the G-optimality criterion of optimality

$$\max_{x \in [a,b]} f^T(x) M^{-1}(x_1, \ldots, x_N) f(x)$$

The minimization of this expression with respect to $x_1, \ldots, x_N$ gives a design which is good not only for the precision of the response function, but also for its robustness with respect to outliers.

Conclusion Even in such a simple model as a linear regression on a real line we found three kinds of "strange" design points.

Acknowledgement The paper was prepared upon the invitation of the ODAM 2011 conference and was also supported by the VEGA grant No. 1/0077/09.

References

[1] Fišerová, E., Kubáček, L., Kunderová, P.: Linear Statistical Models: Regularity and Singularities. *Academia*, Praha, 2007.

[2] Harville, D. A.: Matrix Algebra from a Statistician's Perspective. *Springer*, New York, 1997.

[3] Näther, W.: *Exact designs for regression models with correlated errors.* Statistics **16** (1985), 479–484.

[4] Pázman, A.: Foundations of Optimum Experimentsl Design. *Kluwer*, Dordrecht, 1986.

[5] Pázman, A.: *Information contained in design points of experiments with correlated observations.* Kybernetika **46** (2010), 769–781.

[6] Pázman, A., Pronzato, L.: *On the irregular behavior of LS estimators for asymptotically singular designs.* Statistics and Probability Letters **76** (2006), 1089–1096.

[7] Pukelsheim, F.: Optimal Design of Experiments. *Wiley*, New York, 1993.

[8] Zvára, K.: Regresní analýza. *Academia*, Praha, 1989.