The Gauss Markov Property for the Median

Gilbert W. Bassett Jr.[a]

[a]Department of Economics (m/c144), University of Illinois at Chicago, Box 4348, Chicago IL, 606080, Email u09006@uicvm.cc.uic.edu

Abstract

The median is the best estimator in a class of median-unbiased estimators, just as the sample mean is best in a class of mean-unbiased estimators. The Gauss Markov property for the median, like the property for the mean, reflects restrictions on alternatives, rather than optimality of the estimator. The result is easily accessible and helps to clarify the interpretation of the Gauss Markov Theorem as a supposed justification for the sample mean.

1. INTRODUCTION

The sample mean has the smallest variance in the class of linear and unbiased estimators. This is the Gauss Markov property for the mean. The result holds for a sample, $X_1,\ldots,X_n$, where the X_i's are independent and have a common, but arbitrary, distribution function F_X. (The theorem also holds for the linear model version of the sample mean, the least squares estimator, and it generalizes to non-iid situations, where it known as Aitken's theorem. In this paper attention is restricted to the iid case). Since the property holds for any F_X and any sample size, it has the appearance of a very powerful optimality property.

The sample mean is best for any F_X, but only relative to a class of estimators that does not contain any reasonable alternatives. For example, linear and unbiased estimators include sample means constructed with proper subsets of the data. The mean of half the data is linear, unbiased and--not unexpectedly--inferior to the mean based on all the data. Other linear, unbiased estimators correspond to unequally weighted linear combinations of observations. Again, it is not surprising that the smallest variance estimator within this class is the one that puts equal weight on each of the independent and identically distributed observations.

The Gauss Markov theorem is best viewed as an exercise in identifying restrictions on alternative estimators that are sufficient to make the mean best for all F_X. Since a broad range of situations are covered by the Theorem it is not surprising that the class of alternative estimators needs to be severely restricted. The real purpose is not to establish optimality of the sample mean, but to find restrictions that eliminate all estimators that are superior to the mean.

This nonstandard interpretation of the theorem is reinforced by doing the same exercise, but with a different estimator. The Gauss Markov property for the median, for

example, would identify a class of estimators in which the median is optimal for any F_X. This class would be the analogue of linear and unbiased estimators that yields the sample mean version of the property. Establishing the result would eliminate the sample mean as the only estimator with a Gauss Markov property. With two "best" estimators it would be harder to interpret the Gauss Markov theorem as a genuine optimality result.

The starting point for considering the median version of the Gauss Markov property is an unbiasedness result. The mean is mean-unbiased, meaning that its expectation is equal to the population expectation, for any F_X. The same result holds for the median, except that the definition of mean-unbiased is replaced by median-unbiased. An estimator is median-unbiased if its sampling distribution has a median equal to the population median; that is, if it is 50-50 that the estimate will be above (or below) the population median; see Lehmann(1951). The concept of median-unbiasedness was first considered in Brown(1947). He introduced median-unbiasedness because it accomplishes "as much as the mean-unbiased requirement and has the additional property that it is invariant under one-to-one transformations" (p.583). For example, a mean-unbiased estimate for the variance does not yield--on taking the square root--an unbiased estimate of the standard deviation. Median unbiasedness however is preserved under such a nonlinear transformation. It turns out that the sample median is median-unbiased for any F_X just as the sample mean is mean-unbiased for any F_X. This suggests looking at the class of median-unbiased estimators for alternatives to the sample median.

The sample mean version of the Gauss Markov property imposes the restriction to linear estimators. This class can be conveniently identified by

$$L=\{b|\ \textstyle\sum a_i(X_i-b)=c\} \tag{1.1}$$

where $a=(a_1,\ldots,a_n)$ is nonstochastic and c is a constant. With this representation the set of linear and unbiased estimators, call it LU, is identified by setting $c=0$,

$$LU=\{b|\ \textstyle\sum a_i(X_i-b)=0\} \tag{1.2}$$

that is, $b\epsilon LU$ is linear and mean-unbiased for each vector a ($\Sigma a_i \neq 0$), and conversely. The best estimator in LU (measured by variance) is the one that puts equal weight, $a_i=k$, on each observation. This estimator, call it b*, is the solution to the first-order condition for the equally weighted least squares problem, which means (as can be also seen directly) that b* is the sample mean. This is the Gauss Markov property for the sample mean.

The L estimators in (1.1) are defined by setting the derivative of a weighted sum of squares function equal to a constant. A different class of estimators, which mimics many of the same properties, comes from considering the weighted sum of absolute errors, $\sum a_i|X_i-b|$. This leads to an L* given by,

$$L^* = \{b|\ -\textstyle\sum a_i sgn^*(X_i-b;1) \leq c \leq -\sum a_i sgn^*(X_i-b;-1)\} \tag{1.3}$$

where $sgn^*(x;v) = sgn(x)$ for $x \neq 0$ and sgn(v) otherwise, the a_i's are nonstochastic, and c is a constant. The estimators in both (1.1) and (1.3) are defined by the way they map

observations to estimates, rather than by any statistical property[1]. Mean-unbiased estimators in L are obtained by setting c=0. In a similar fashion it is shown below that median-unbiased estimators, L*U*, are obtained by setting c=0 in (1.3).

Let b* in L*U* denote the estimator with $a_i=k$, a constant. This is the solution of a first-order condition, now the least absolute error problem, and b* is therefore the sample median. The median is usually defined and thought of in terms of an ordered sample. It is often useful however, just as with the least squares definition of the mean, to define the median as well as the other quantiles by a minimization problem; see, for example, Koenker and Bassett(1978) and Bassett and Koenker(1978 and 1982). The question is whether b* the best estimator in L*U*.

A ranking that is convenient for assessing the median is based on a criterion considered by Pitman (1938). Let ß denote the median of X and let the best estimator in L*U* be the one whose sampling distribution is most concentrated about ß; that is, a $b_o \epsilon L^*U^*$ such that,

$$P(|b_o-ß| < w) \geq Pr(|b-ß| < w) \qquad (1.4)$$

for all $w > 0$ and all other $b \epsilon L^*U^*$. This expresses the fact that b_o is most concentrated around the population median. It is shown below that this best estimator is the sample median.

The results in section 2 are derived under the assumption that F_X is continuous and increasing. The population median is then a point, not an interval, and proofs are simpler than the general case; modifications that are required for the general case are omitted. A summary is in section 3.

2. THE BLUE ESTIMATOR

Notation and Definitions. Let $X_1,\ldots,X_n$ be i.i.d. with distribution F_X. F_X is assumed to be continuous and increasing, but otherwise arbitrary. The median of X is ß where $F_X(ß)=½$.

Let $\rho(b:a)$ denote $\sum a_i|X_i-b|$, where $a \epsilon R^n$ is nonstochastic. The right and left derivatives of ρ are given by

---

[1]It is important distinguish between (i) minimization problems that identify estimates and (ii) statistical criteria used to evaluate estimators. The sample mean estimate, for example, can be identified in a number of alternative ways, including as the estimate that minimizes the sum of squared errors. This does not imply that sample mean is the minimum expected squared error estimator. The minimum expected squared error estimator depends on the population distribution. The sample mean is best when sampling from a Gaussian distribution, but it is inferior to the sample median when sampling from a Laplace distribution even under the expected squared error criterion. (Both estimates are mean unbiased and the median has a smaller variance). That the best estimator depends on the shape of the population distribution is the reason for introducing restrictions on the set of alternative estimators when devising Gauss Markov properties.

$\psi_+(b:a) = -\sum a_i \text{sgn}^*(X_i-b;-1)$
$\psi_-(b:a) = -\sum a_i \text{sgn}^*(X_i-b;1)$

Let b(a) denote the minimum set of ρ. Since ρ is convex the first order condition for a minimum is,

$$b(a)=\{b \mid \psi_-(b:a) \le 0 \le \psi_+(b:a)\} \tag{2.1}$$

Let L*U* denote all b(a) where a satisfies the following conditions:

$$a_i \ge 0,\ i=1,\ldots,n \tag{2.2}$$
$$\Sigma a_i s_i \neq 0 \text{ for all } s \epsilon S. \tag{2.3}$$

where S is the vectors s in R^n with components $s_i = \pm 1$.

The first condition (2.2) makes ρ convex for all $x \epsilon R^n$. The second condition insures that the minimum set of ρ is a point, not an interval, and hence that b(a) is a well-defined estimator. (The minimum set of ρ is a point when the right derivative is positive and the left derivative is negative. Condition (2.3) guarantees that this will be the case for all $x \epsilon R^n$).

Remark. The uniqueness condition (2.3) is not restrictive. It can be always insured with a tie-breaking rule that involves a slight perturbation of an a_i. For example, let $a_1=1$ when $\Sigma a_i s_i \neq 0$ and $a_1 = 1+\delta$, otherwise, where $\delta > 0$ is very small. This perturbation does not alter the estimate when b(a) is a point. If (2.1) is a nondegenerate interval, this slight perturbation causes b(a) to be the element in (2.1) closest to the first observation.

Definition of the median. The median will be denoted by b* and identified as the b(a) with all the a_i's equal to one another, or nearly so. That is, the median is defined by setting $a=a^*$ where $a^*=(1+\delta,1,\ldots,1)$ where δ is small. This corresponds to the usual definition of the median when n is odd. When n is even it corresponds to a tie-breaking rule in which "the" median is the endpoint of the median interval closest to the first observation.

Remarks: (1). The element closest to the i^{th} observation would be selected by adding δ to a_i instead of a_1. Any such perturbation of a_i would achieve a unique estimate and lead to the same distribution for b*.

(2). An advantage of the tie-breaking convention adopted here is that the median stays median-unbiased when n is even. This would not necessarily be the case with other types of tie-breaking rules; that is, rules that did not involve a slight perturbing of the a_i's. For a simple example, if n=2 and the median is defined as the midpoint of the median interval, also equal to the sample mean, then the estimate (viz., mean) will not be median-unbiased when F_X is asymmetric.

Examples. If a=(1,1,1) then b(a)=b* is the median. When a=(1,1,1,1) condition (2.3) is violated, but a unique estimate is defined by $a^*=(1+\delta,1,1,1)$; at x=(0,2,3,9) the b* estimate is 2, the endpoint of the median interval [2,3] that is closest to 0--the first observation.

The distribution of b(a)ϵL*U*. General results for b(a) estimators are greatly simplified by the following monotonicity property, best known for the median:

$$F_X(b(X_1,\ldots,X_n)) = b(F_X(X_1),\ldots,F_X(X_n)) = b(V_1,\ldots,V_n)$$

where $F_X(X_i)=V_i$ is a uniform random variable on [0,1]. (This equivariance property for b(a) estimators can be easily verified on noting, $\text{sgn}^*(X_i-b;\pm 1) = \text{sgn}^*(F_X(X_i)-F_X(b);\pm 1)$). This says that the distribution of $F_X(b(X))$ is the same as the distribution of b(V), and hence, without loss of generality, we can restrict attention to a sample of uniformly distributed random variables.

Consider the event, $[F_X(b(X)<d]=[b(V)<d]$. This says the convex function ρ attains its minimum to the left of d, or, since ρ is convex, the left derivative at d is positive,
$P[F_X(b)<d] = P[-\sum a_i \text{sgn}^*(V_i-d;1)>0]$.
Since $\text{sgn}^*(V_i-d;1)$ takes values -1 and +1 with probability d and 1-d this can be written as

$$P[F_X(b)<d] = P[-\sum a_i S_i(d)>0].$$

where $S_i(d)$ is -1 and +1 with probability d and 1-d, respectively.

The proof of median-unbiasedness now follows directly from consideration of the simple random variables S_i.

Theorem 2.1. Each $b\epsilon L^*U^*$ is median-unbiased.

Proof: b(a) is median-unbiased if $P[F_X(b(a))<½]=½$, which follows if $F_X(b(a))$ is symmetric about ½. This can be verified directly using the fact that $-S_i(½+d)$ is the same as $S_i(½-d)$. The steps are,
$P[F_X(b(a))<½+d] = P[-\sum a_i S_i(½+d)>0] = 1 - P[-\sum a_i S_i(½+d)\le 0] =$
$1 - P[-\sum a_i S_i(½+d)<0] = 1 - P[\sum a_i S_i(½-d)<0] = 1 - P[-\sum a_i S_i(½-d)>0] =$
$1 - P[F_X(b(a))<½-d] = P[F_X(b(a))\ge ½-d]$,
which means $F_X(b(a))$ is symmetric about ½.

The proof that b* is the best estimator also follows directly from properties of the simple S_i random variables.

Theorem 2.2. The best estimator in the L*U* class of median-unbiased estimators is the median.

Proof: We have to show that for each $b=b(a)\epsilon L^*U^*$ and each $w>0$, $P(|b^*-\beta|<w) \ge Pr(|b-\beta|<w)$, or $P(½-d'<F_X(b^*)<½+d) > P(½-d'<F_X(b)<½+d)$ where d and d′ are in (0,½); $d = F_X(\beta+w)-½$ and $-d' = F_X(\beta-w)-½$. We will show that $P[F_X(b^*)<½+d] > P[F_X(b)<½+d]$; the similar steps used for $P[F_X(b^*)<½-d'] < P[F_X(b)<½-d']$ are omitted.

Let $b\epsilon L^*U^*$ be arbitrary, fix $d>0$, and consider, $P[F_X(b)<½+d] =$
$P[-\sum a_i S_i(½+d)>0]$. This probability can be expressed as the sum of the probabilities at each of the $s\epsilon S$ where $-\Sigma a_i s_i>0$, or

$$P[F_X(b)<½+d]=\sum_{s\epsilon S} I[-\sum_i a_i s_i>0]\, Pr[S(½+d)=s] \qquad (2.4)$$

where I[ ] is the (1 or 0) indicator function.

Write the indicator as the sum of two parts (recall that $\Sigma a_i^* s_i \neq 0$),
$I(-\Sigma a_i s_i>0) = I[-\Sigma a_i s_i>0, -\Sigma a_i^* s_i>0] + I[-\Sigma a_i s_i>0, -\Sigma a_i^* s_i<0]$
and substitute into the right hand side of (2.4)

$$\sum_{s\in S} I[-\Sigma a_i s_i>0,\ -\Sigma a_i * s_i>0]Pr[S(½+d)=s] +$$
$$\sum_{s\in S} I[-\Sigma a_i s_i>0,\ -\Sigma a_i * s_i<0]Pr[S(½+d)=s]$$

Change variables in the latter term from s to -s

$$\sum_{s\in S} I[\Sigma a_i s_i>0,\ \Sigma a_i * s_i<0]Pr[S(½+d)=-s]$$

and multiply inside the indicator by -1,

$$\sum_{s\in S} I[-\Sigma a_i s_i<0,\ -\Sigma a_i * s_i>0]Pr[S(½+d)=-s]$$

Now, $-\Sigma a_i^* s_i>0$ implies, $P[S(½+d)=-s] < P[S(½+d)=s]$ (see appendix) so,

$$P[F_X(b)<½+d] \leq$$
$$\sum_{s\in S} I[-\Sigma a_i s_i>0,\ -\Sigma a_i * s_i>0]P[S(½+d)=s] +$$
$$\sum_{s\in S} I[-\Sigma a_i s_i<0,\ -\Sigma a_i * s_i>0]P[S(½+d)=s]$$

and with $\Sigma a_i s_i \neq 0$ the two terms combine into

$$\sum_{s\in S} I[-\Sigma a_i * s_i>0]P[S(½+d)=s]$$

But this is the distribution of b at $a=a^*$, the median, and hence proves the theorem.

3. CONCLUSION

The Gauss Markov property for the median shows that the median is the best estimator in a class of median-unbiased estimators. The result can be readily proved without advanced methods and it can be used to introduce alternative notions of unbiasedness and dispersion. The parallels between the Gauss Markov properties for the mean and median are summarized in Table 1.

The Gauss Markov property does not imply that the median is better than the mean. It instead highlights the fact that the appearance of optimality in the statement of the Theorem really comes from a judicious choice of alternatives. Anything does looks good in the right light. This helps clarify the interpretation of the Gauss Markov Theorem for the sample mean[1].

---

[1]I would like to acknowledge comments from Bill Farebrother, Erich Lehmann, and Stephen Stigler Stigler on an earlier version of this paper. The usual disclaimer applies.

Table 1
The Gauss Markov Property for the Mean and Median

Model	$X_i=\beta+e_i$, iid, $E(e_i)=0$	$X_i=\beta+e_i$, iid, $M(e_i)=0$
"Linear"	$\{b \mid \sum a_i(X_i-b)=c\}$	$\{b \mid \sum -a_i sgn^*(X_i-b) \approx c\}^a$
"Unbiased"	$E(b)=\beta$	$M(b)=\beta$
LU	$\{b \mid \sum a_i(X_i-b)=0\}$	$\{b \mid \sum -a_i sgn^*(X_i-b) \approx 0\}^a$
M-estimate	$\min \sum a_i(X_i-b)^2$	$\min \sum a_i \lvert X_i-b \rvert$
Dispersion	Variance	Pitman
BLUE	mean	median

[a]See (1.3)

4. APPENDIX

Result: If $d>0$ and $-\Sigma a_i^* s_i>0$ then $P[S(\frac{1}{2}+d)=s] > P[S(\frac{1}{2}+d)=-s]$.

Proof: The components of S are iid and the probability of, $S(\frac{1}{2}+d)=s$, depends on the number of components in s that are -1 and +1. This probability is

$$P[S(\tfrac{1}{2}+d)=s]=(\tfrac{1}{2}+d)^{\Sigma\frac{1}{2}(1-s_i)}(\tfrac{1}{2}-d)^{\Sigma\frac{1}{2}(1+s_i)}$$

$$=[(\tfrac{1}{2}+d)(\tfrac{1}{2}-d)]^{\frac{n}{2}}\left(\frac{\frac{1}{2}+d}{\frac{1}{2}-d}\right)^{-\frac{1}{2}\Sigma s_i}$$

$$=K(d)\ M(d)^{-\frac{1}{2}\Sigma s_i}$$

where $K(d)=[(\frac{1}{2}+d)(\frac{1}{2}-d)]^{n/2}$ and $M(d)=[(\frac{1}{2}+d)/(\frac{1}{2}-d)]>1$. Since $-\Sigma a_i^* s_i>0$ implies $-\Sigma s_i>0$ we have

$$>K_n(d)\ M(d)^{\frac{1}{2}\Sigma s_i}$$

$$=P[S(\tfrac{1}{2}+d)=-s].$$

This result just says that if -1 is more likely than +1 then S=s is more likely than S=-s whenever the sum of the s_i's is negative--because then there are more -1's in s.

5. REFERENCES

Bassett, G.W. and R.W.Koenker(1978). The Asymptotic Theory of Least Absolute Error Regression, *Journal of the American Statistical Association*, Vol.73, No. 363, September 1978, 618-622.

Bassett, G.W. and R.W.Koenker(1982). An Empirical Quantile Function for Linear Models with i.i.d. Errors, *Journal of the American Statistical Association*, Vol. 77, No. 378 June 1982.

Brown, George W. (1947). On Small-Sample Estimation, *Annals of Mathematical Statistics*, Vol. 18 p. 582-585.

Koenker,R.W. and G.W.Bassett(1978). Regression Quantiles, *Econometrica*, Vol. 46, No. 1, January 1978, 33-50.

Lehmann, E.L. (1951). A General Concept of Unbiasedness, *Annals of Mathematical Statistics*, Vol. 22 p. 587-592.

Pitman, E.J.G. (1938). The Estimation of the Location and Scale Parameters of a Continuous Population of any Given Form, *Biometrika* p.391-421.