

This document is an update to Obenchain RL (1975) *Journal of the American Statistical Association* 70: 375-379. This update attempts to clarify some points by increasing the length of the published version by about 10% and also fixing some typos.

Residual Optimality: Ordinary Vs. Weighted Vs. Biased Least Squares

Bob Obenchain

In the general linear model with observations not necessarily uncorrelated or homoscedastic, Gauss-Markov regression coefficients are superior to ordinary unweighted least squares in the well known BLU sense when the assumed model is correct. However, it is shown that there is a weaker, but always applicable, minimum overall mean squared error sense in which Gauss-Markov residuals and biased residuals are inferior to ordinary least squares residuals as estimators of possible lack-of-fit in the model. This optimality of ordinary least squares is further illustrated by three other types of results about residuals.

1. INTRODUCTION

These notes give formal, mathematical properties of residuals corresponding to ordinary least squares or diagonally weighted least squares estimators that make them superior to the residuals from Gauss-Markov and biased linear estimators when the problem is to estimate possible lack-of-fit in the assumed expectation model. These formal properties insure the usefulness of ordinary or diagonally weighted least squares residuals in informal, exploratory “diagnostic procedures” such as those of Draper and Smith [4, Ch. 3], Wilk and Gnanadesikan [12] and Larsen and McCleary [8].

The general linear model and all basic notation are presented in Section 2. A coefficient of residual variation, $CV(\mathbf{r})$, is defined in Section 3, and all unbiased linear estimators are shown to achieve the same $CV(\mathbf{r})$ value, while biased estimators can

achieve smaller values. In Section 4 it is shown that, among all linear estimators, ordinary least squares produces residuals which are as small in expected size, variability, and mean squared error risk as is possible in an overall sense which applies even when the assumed model is incorrect. A measure \hat{R} of the possible pathology of weighted least squares residuals corresponding to the pairing of an assumed error covariance structure, \mathbf{V} , with an assumed design or regressor matrix, \mathbf{X} , is considered in Section 5. Transformations of the general linear model to create homoscedastic observations are considered in Section 6; although this commonly seen motivation for non-diagonally weighted least squares is rejected because it destroys the labels on (identity of) residuals, another transformational invariance argument is shown to motivate diagonally weighted least squares. The appendix is a tabular summary.

2. NOTATION

Let \mathbf{y} denote a realization of an $n \times 1$ vector of random variables, \mathbf{Y} , and let \mathbf{X} denote an $n \times p$ matrix of rank $p \leq n$ containing known design or regressor values. Let \mathbf{m} denote the unknown conditional expectation of \mathbf{Y} given \mathbf{X} , $E(\mathbf{Y}|\mathbf{X}) = \mathbf{m}$, and let \mathbf{S} denote the unknown conditional dispersion (variance-covariance) matrix $D(\mathbf{Y}|\mathbf{X}) = \mathbf{S}$. On the basis of assumption or partial information, consider the following general linear model. Let the expectation model be $\mathbf{m} = \mathbf{X}\mathbf{b}$, where \mathbf{b} is a $p \times 1$ vector of unknown parameters (regression coefficients.) The expectation model will be said to be correct when $\mathbf{m} \in C(\mathbf{X})$, the column space of \mathbf{X} (i.e. $\mathbf{m} = \mathbf{X}\mathbf{b}$ for some \mathbf{b} .) Let the dispersion model be $\mathbf{S} = \mathbf{s}^2\mathbf{V}$, where \mathbf{V} is the assumed $n \times n$ positive definite covariance structure, and \mathbf{s}^2 is an unknown variance multiplier. The dispersion model will be said to be correct when \mathbf{S} actually is proportional to the assumed \mathbf{V} . The resulting complete model is ($\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, $E(\mathbf{e}|\mathbf{X}) = \mathbf{0}$, and $D(\mathbf{e}|\mathbf{X}) = \mathbf{s}^2\mathbf{V}$), where \mathbf{e} is an unobservable $n \times 1$ vector of error random variables. In the following, it will be convenient not to distinguish in the notation between \mathbf{y} and \mathbf{Y} and, therefore, between estimates and estimators.

A linear estimator of \mathbf{b} is any statistic of the form $\hat{\mathbf{b}} = \mathbf{L}\mathbf{y}$, where \mathbf{L} is a fixed $p \times n$ matrix that can depend upon \mathbf{X} and \mathbf{V} but not upon \mathbf{y} . The condition for $\hat{\mathbf{b}}$ to be unbiased when $\mathbf{m} = \mathbf{X}\mathbf{b}$ for any \mathbf{b} is that $\mathbf{L}\mathbf{X} = \mathbf{I}$. Since $\text{rank}(\mathbf{X}) = p$ by assumption, $\mathbf{L} = \mathbf{X}^-$ is then any g-inverse of \mathbf{X} , [10, p. 24].

Defining $\mathbf{Q} \equiv \mathbf{I} - \mathbf{X}\mathbf{L}$, the residual vector corresponding to $\hat{\mathbf{b}} = \mathbf{L}\mathbf{y}$ is $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = \mathbf{Q}\mathbf{y}$. It follows that $E(\mathbf{r}|\mathbf{X}) = \mathbf{0}$ if $\mathbf{m} \in C(\mathbf{X})$ and $\mathbf{L} = \mathbf{X}^-$. Similarly, the relationships $D(\hat{\mathbf{b}}|\mathbf{X}) = \mathbf{L}\mathbf{S}\mathbf{L}'$ and $D(\mathbf{r}|\mathbf{X}) = \mathbf{Q}\mathbf{S}\mathbf{Q}'$ hold in general.

The *lack-of-fit* in $\mathbf{m} = \mathbf{X}\mathbf{b}$ which can be estimated using $\hat{\mathbf{b}} = \mathbf{L}\mathbf{y}$ is defined to be $E(\mathbf{r}|\mathbf{X}) = \mathbf{Q}\mathbf{m}$. If $\mathbf{m} \in C(\mathbf{X})$ and $\mathbf{L} = \mathbf{X}^-$, this lack-of-fit would, of course, be zero. When $\mathbf{m} = \mathbf{X}\mathbf{b}$ for some \mathbf{b} and $\hat{\mathbf{b}}$ is unbiased, the residual vector \mathbf{r} is useful in describing the unobserved vector of errors, \mathbf{e} , and in defining a residual quadratic form estimator, \mathbf{s}^2 , of \mathbf{s}^2 .

The ordinary unweighted least squares estimator of \mathbf{b} will be denoted by $\mathbf{b}^0 = \mathbf{L}^0\mathbf{y}$, where $\mathbf{L}^0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}^+$ and a $+$ superscript denotes the unique Moore-Penrose inverse of a matrix, [10, p. 26]. The corresponding residual vector will be denoted by $\mathbf{r}^0 = \mathbf{Q}^0\mathbf{y}$ where $\mathbf{Q}^0 = \mathbf{I} - \mathbf{X}\mathbf{X}^+$. Note that \mathbf{Q}^0 is the uniquely determined, symmetric and idempotent projection matrix for the space of all $n \times 1$ vectors orthogonal to $C(\mathbf{X})$.

The optimally weighted, Gauss-Markov estimator of \mathbf{b} [1] will be denoted by $\hat{\mathbf{b}} = \hat{\mathbf{L}}\mathbf{y}$, where $\hat{\mathbf{L}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. The corresponding residual vector is $\hat{\mathbf{r}} = \hat{\mathbf{Q}}\mathbf{y}$ for $\hat{\mathbf{Q}} = \mathbf{I} - \mathbf{X}\hat{\mathbf{L}}$.

This manuscript describes senses in which \mathbf{b}^0 is superior to both $\hat{\mathbf{b}}$ (when $\hat{\mathbf{b}} \neq \mathbf{b}^0$) and to biased linear estimators such as $\mathbf{b}^* = (\mathbf{X}'\mathbf{X} + \mathbf{K})^{-1}\mathbf{X}'\mathbf{y}$ when $\mathbf{K} \neq \mathbf{0}$; see [7] and [9]. These new results will thus stand in marked contrast to well known senses in which $\hat{\mathbf{b}}$ and/or \mathbf{b}^* are superior to \mathbf{b}^0 . Specifically, the BLU (minimum variance or “best” linear unbiased) estimator of \mathbf{b} is $\hat{\mathbf{b}}$ when the expectation and dispersion models are both correct [1]. Since $\mathbf{r} - \mathbf{e} = \mathbf{X}(\mathbf{b} - \mathbf{b})$ in this case, it follows that $E(\mathbf{r} - \mathbf{e} | \mathbf{X}) = \mathbf{0}$ when $\mathbf{L} = \mathbf{X}^-$ and that the difference $D(\mathbf{r} - \mathbf{e} | \mathbf{X}) = \mathbf{X}D(\mathbf{b}|\mathbf{X})\mathbf{X}'$ minus $D(\hat{\mathbf{r}} - \mathbf{e} | \mathbf{X})$ is nonnegative definite. Thus, among all $\mathbf{r} = (\mathbf{I} - \mathbf{X}\mathbf{X}^-)\mathbf{y}$, $\hat{\mathbf{r}}$ deviates from \mathbf{e} with minimum mean squared error when the assumed expectation and variance models are

both correct. Similarly, with regard to optimal biased linear estimation, Hoerl and Kennard [7] show that, when $\mathbf{S} = \mathbf{s}^2\mathbf{I}$, $\mathbf{K} = k\mathbf{I}$, and $\mathbf{b}\mathbf{c}\mathbf{b}$ is bounded, there exists an unknown $k > 0$ such that \mathbf{b}^* has smaller summed mean squared error, $E[(\mathbf{b} - \mathbf{b}')(\mathbf{b} - \mathbf{b}') | \mathbf{X}]$, than $\mathbf{b}^0 = \hat{\mathbf{b}}$.

3. A COEFFICIENT OF RESIDUAL VARIATION

A scalar valued coefficient of multivariate variation (a generalized ratio of standard error to expected value) can be defined for \mathbf{r} using the formulas $E(\mathbf{r}|\mathbf{X}) = \mathbf{Q}\mathbf{m}$ and $D(\mathbf{r}|\mathbf{X}) = \mathbf{Q}\mathbf{S}\mathbf{Q}'$ which always apply (even when the assumed models are incorrect.) Thus we define $CV(\mathbf{r}) \equiv +\infty$ when $E(\mathbf{r}|\mathbf{X}) = 0$ and otherwise utilize the formula

$$CV(\mathbf{r}) = [\mathbf{m}'\mathbf{Q}'(\mathbf{Q}\mathbf{S}\mathbf{Q}')^{-1}\mathbf{Q}\mathbf{m}]^{-1/2} \quad (3.1)$$

In the proof of the following theorem, it will become clear that all g-inverses in (3.1) yield the same numerical value for $CV(\mathbf{r})$.

Theorem 1: If \mathbf{b} is linear ($\mathbf{b} = \mathbf{L}\mathbf{y}$), then $CV(\mathbf{r}) \leq CV(\mathbf{r}^0)$. When \mathbf{b} is unbiased ($\mathbf{L} = \mathbf{X}$), $CV(\mathbf{r}) = CV(\mathbf{r}^0)$. On the other hand, when \mathbf{b}^* is biased ($\mathbf{L}^* \neq \mathbf{X}$), $CV(\mathbf{r}^*) < +\infty$ for some $\mathbf{m} \in C(\mathbf{X})$. Finally, $CV(\mathbf{r}^*) < CV(\mathbf{r}^0) < +\infty$ for some \mathbf{m} with nonzero components in both $C(\mathbf{X})$ and its orthogonal complement.

Proof: $CV(\mathbf{r}) = (\mathbf{m}'\mathbf{S}^{-1/2}\mathbf{P}\mathbf{S}^{-1/2}\mathbf{m})^{-1/2}$ where $\mathbf{P} = \mathbf{S}^{1/2}\mathbf{Q}'(\mathbf{Q}\mathbf{S}\mathbf{Q}')^{-1}\mathbf{Q}\mathbf{S}^{1/2}$ is the uniquely determined orthogonal projection matrix for the column space $C(\mathbf{S}^{1/2}\mathbf{Q}')$ [10, p. 47, (vi)]. Thus, $CV(\mathbf{r})$ is the reciprocal of the length of the projection of $\mathbf{S}^{-1/2}\mathbf{m}$ onto $C(\mathbf{S}^{1/2}\mathbf{Q}')$. Since $\mathbf{Q}^0\mathbf{Q} = \mathbf{Q}^0 = \mathbf{Q}^0'$ for every $\mathbf{Q} = \mathbf{I} - \mathbf{X}\mathbf{L}$, it follows that $\mathbf{S}^{1/2}\mathbf{Q}^0\mathbf{a} = \mathbf{S}^{1/2}\mathbf{Q}'\mathbf{Q}^0\mathbf{a}$ for every \mathbf{a} . Thus $C(\mathbf{S}^{1/2}\mathbf{Q}')$

always contains $C(\mathbf{S}^{1/2}\mathbf{Q}^0)$, and therefore $CV(\mathbf{r}) \leq CV(\mathbf{r}^0)$.

Now if $\mathbf{Q} = \mathbf{I} - \mathbf{X}\mathbf{X}$, then $\mathbf{Q}\mathbf{Q}^0 = \mathbf{Q}$ also holds. Thus $C(\mathbf{S}^{1/2}\mathbf{Q}') = C(\mathbf{S}^{1/2}\mathbf{Q}^0)$ in this case, and $CV(\mathbf{r}) = CV(\mathbf{r}^0)$.

Finally, note that $(\mathbf{Q}\mathbf{S}^{1/2})(\mathbf{S}^{-1/2}\mathbf{m}) = \mathbf{Q}\mathbf{X}\mathbf{b}$ when $\mathbf{m} \in C(\mathbf{X})$. It follows that $\mathbf{Q}\mathbf{X} = \mathbf{0}$ and $CV(\mathbf{r}) = +\infty$ for every \mathbf{b} iff $\mathbf{L} = \mathbf{X}$ (see [10, p. 24]). Thus, $CV(\mathbf{r}^*) < +\infty$ for some $\mathbf{m} \in C(\mathbf{X})$ when $\mathbf{L}^* \neq \mathbf{X}$. In other words, $C(\mathbf{S}^{1/2}\mathbf{Q}^*)$ is strictly larger than $C(\mathbf{S}^{1/2}\mathbf{Q}^0)$ because it contains vectors not orthogonal to $C(\mathbf{S}^{-1/2}\mathbf{X})$. It follows that $CV(\mathbf{r}^*) < CV(\mathbf{r}^0) < +\infty$ for some $\mathbf{m} \notin C(\mathbf{X})$ such that $\mathbf{m}\mathbf{c}\mathbf{X} \neq \mathbf{0}\mathbf{c}$.

Theorem 1 has at least one interesting implication. $CV[(\mathbf{I} - \mathbf{X}\mathbf{X})\mathbf{y}] = CV(\mathbf{r}^0)$ implies that, corresponding to any sense in which $E[(\mathbf{I} - \mathbf{X}\mathbf{X})\mathbf{y} | \mathbf{X}] \geq E(\mathbf{r}^0|\mathbf{X})$, there must also be some sense in which $D[(\mathbf{I} - \mathbf{X}\mathbf{X})\mathbf{y} | \mathbf{X}] \geq D(\mathbf{r}^0|\mathbf{X})$. A result of this type, except that unbiasedness is not needed, is given in Theorem 2. Thus $CV(\mathbf{r}^*) \leq CV(\mathbf{r}^0)$ and $D(\mathbf{r}^*|\mathbf{X}) \geq D(\mathbf{r}^0|\mathbf{X})$ implies that, even relative to their larger dispersion, biased residuals have larger means than unbiased residuals.

It is also clear from the proof of Theorem 1 that the residual quadratic form estimator $\mathbf{s}^2 = \mathbf{r}'(\mathbf{Q}\mathbf{V}\mathbf{Q}')^{-1}\mathbf{r} / (n-p)$ of \mathbf{s}^2 is uniquely determined for every $\mathbf{L} = \mathbf{X}$. The most commonly seen formula for \mathbf{s}^2 is $\mathbf{s}^2 = \hat{\mathbf{r}}'\mathbf{V}^{-1}\hat{\mathbf{r}} / (n-p)$, which results from the preceding general formula because of the special form of $\hat{\mathbf{Q}} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$.

4. RESIDUAL OPTIMALITY OF ORDINARY LEAST SQUARES

Note that $\mathbf{S} = \mathbf{r}\mathbf{r}'$ is a rank one estimator of the residual mean squared error matrix $E(\mathbf{r}\mathbf{r}'|\mathbf{X})$.

It follows that $\mathbf{Q} = \mathbf{Q}^\circ$ leads to the smallest possible such \mathbf{S} estimator in the sense that all of the eigen values of \mathbf{S}° [namely $\mathbf{r}^{\circ'}\mathbf{r}^\circ$ and $(n-1)$ zeros] are simultaneously minimized. Thus, the following results on the minimization of $E(\mathbf{r}|\mathbf{X})$, $E(\mathbf{r}\mathbf{r}'|\mathbf{X})$, and $D(\mathbf{r}|\mathbf{X})$ when $\mathbf{Q} = \mathbf{Q}^\circ$ may not be surprising.

Theorem 2: Among all residual vectors $\mathbf{r} = \mathbf{Q}\mathbf{y}$ corresponding to linear estimators $\mathbf{b} = \mathbf{L}\mathbf{y}$ for the general linear model ($\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, $E(\mathbf{e}|\mathbf{X}) = \mathbf{0}$, and $D(\mathbf{e}|\mathbf{X}) = \mathbf{s}^2\mathbf{V}$), ordinary unweighted least squares ($\mathbf{Q}^\circ = \mathbf{I} - \mathbf{X}\mathbf{X}^+$) simultaneously minimizes (i) $E(\mathbf{r}|\mathbf{X})'E(\mathbf{r}|\mathbf{X}) = \mathbf{m}'\mathbf{Q}'\mathbf{Q}\mathbf{m}$, (ii) all of the eigen values of $D(\mathbf{r}|\mathbf{X}) = \mathbf{Q}\mathbf{S}\mathbf{Q}'$, and (iii) all of the eigen values of $E(\mathbf{r}\mathbf{r}'|\mathbf{X}) = \mathbf{Q}(\mathbf{S} + \mathbf{m}\mathbf{m}')\mathbf{Q}'$, for every true $\mathbf{m} = E(\mathbf{y}|\mathbf{X})$ and for every true $\mathbf{S} = D(\mathbf{y}|\mathbf{X})$.

Proof: Since $E(\mathbf{r}^\circ|\mathbf{X}) = \mathbf{Q}^\circ\mathbf{m}$ is the perpendicular from \mathbf{m} to $C(\mathbf{X})$ [10, p.10, (iv), and p.47], $E(\mathbf{r}^\circ|\mathbf{X})$ is the minimum lack-of-fit of $\mathbf{X}\mathbf{b}$ to \mathbf{m} , and part (i) of Theorem 2 holds even without restricting attention to linear estimates of \mathbf{b} . However, note that parts (i), (ii), and (iii) all follow from the fact that the choice $\mathbf{Q} = \mathbf{Q}^\circ$, simultaneously minimizes all of the eigen values of $\mathbf{Q}\mathbf{A}\mathbf{Q}'$ for any symmetric \mathbf{A} when \mathbf{Q} is restricted to have the form $\mathbf{Q} = \mathbf{I} - \mathbf{X}\mathbf{L}$. To show this, note that \mathbf{Q}° has a (non-unique) decomposition $\mathbf{Q}^\circ = \mathbf{B}\mathbf{B}'$ in which the $n \times (n-p)$ matrix \mathbf{B} is semi-orthogonal, $\mathbf{B}'\mathbf{B} = \mathbf{I}$. Now, since $\mathbf{Q}^\circ\mathbf{Q} = \mathbf{Q}^\circ$ when $\mathbf{Q} = \mathbf{I} - \mathbf{X}\mathbf{L}$, it follows that $\mathbf{Q}^\circ\mathbf{Q}\mathbf{A}\mathbf{Q}'\mathbf{Q}^\circ = \mathbf{Q}^\circ\mathbf{A}\mathbf{Q}^\circ$ and that the $n-p$ largest eigen values $\mathbf{Q}^\circ\mathbf{A}\mathbf{Q}^\circ$ and of $\mathbf{B}'\mathbf{Q}\mathbf{A}\mathbf{Q}'\mathbf{B}$ are equal. The Poincare Separation Theorem [10, p. 64] then shows that each eigen value of $\mathbf{Q}\mathbf{A}\mathbf{Q}'$ cannot be smaller than the corresponding eigen value of $\mathbf{Q}^\circ\mathbf{A}\mathbf{Q}^\circ$.

All common “overall” measures of the size of a dispersion (or mean squared error) matrix, D , are monotone increasing functions of its eigen values. For example, the “total variance,” $\text{trace}(D)$, is the sum of eigen values, and the sum of squares of variances and covariances, $\|D\|^2 = \text{trace}(D^2)$, is the sum of squares of eigen values. Since \mathbf{Q}° simultaneously minimizes all of the eigen values of $D(\mathbf{r}|\mathbf{X})$ and of $E(\mathbf{r}\mathbf{r}'|\mathbf{X})$, Theorem 2 shows that \mathbf{Q}° minimizes all common overall measures of the size of these matrices.

The “strength” of the residual optimality property of \mathbf{b}° is that it is ALWAYS APPLICABLE ($\mathbf{m} =$ or $\neq \mathbf{X}\mathbf{b}$ and $\mathbf{S} =$ or $\neq \mathbf{s}^2\mathbf{V}$.) However, the residual optimality property of \mathbf{b}° is “weaker” in one sense than is the minimum variance property of $\hat{\mathbf{b}}$ when $\mathbf{S} = \mathbf{s}^2\mathbf{V}$ is the correct dispersion model. Specifically, $D(\mathbf{b}^\circ|\mathbf{X}) - D(\hat{\mathbf{b}}|\mathbf{X})$ is a nonnegative definite matrix when $\mathbf{S} = \mathbf{s}^2\mathbf{V}$, but $D(\hat{\mathbf{r}}|\mathbf{X}) - D(\mathbf{r}^\circ|\mathbf{X})$ may not be nonnegative definite. For example, suppose that the i^{th} diagonal element of \mathbf{V} is quite small relative to the other diagonal elements. The i^{th} Gauss-Markov residual will then be relatively small, and its variance $D(\hat{\mathbf{r}}_i|\mathbf{X})$ can be smaller than $D(\mathbf{r}_i^\circ|\mathbf{X})$ when $\mathbf{S} = \mathbf{s}^2\mathbf{V}$ is a correct dispersion model. The optimally weighted least squares regression will indeed provide a relatively close and accurate fit to the i^{th} response, y_i , in these situations.

For completeness, it is remarked that Grossman and Styan [6] have proven a result somewhat similar to part (ii) of Theorem 2 about Theil's BLUS residuals. The main distinctions are that they examine the covariance matrix of the difference between $n-p$ fitted residuals and the corresponding \mathbf{e} 's when $\mathbf{V} = \mathbf{I}$, and their method of proof is quite different from that used here.

It also seems appropriate to give an intuitive explanation of why \mathbf{b}^0 can lead to residuals with minimum overall mean squared error and yet not possess mean squared error optimality properties as an estimator of the regression coefficient vector, \mathbf{b} . Hoerl and Kennard [7] showed that the \mathbf{b}^0 vector is “too long” in their summed mean squared error sense. However, since the OLS residual vector, \mathbf{r}^0 , is already the shortest possible residual vector, it is intuitively clear that no reduction in risk can result from the introduction of residual bias.

5. PATHOLOGICAL \mathbf{V} , \mathbf{X} PAIRINGS

The simple example of Canner [3] illustrates how $\hat{\mathbf{r}}$ can have curious properties when the assumed form of \mathbf{V} is “extreme.” The interplay between \mathbf{V} and \mathbf{X} might be illustrated by a comparison of principal component decompositions for \mathbf{e} and $\hat{\mathbf{r}}$. However, only one summary statistic, \hat{R} , which measures the pathology of \mathbf{V} with respect to \mathbf{X} will be discussed here. Let R denote the multiple correlation between \mathbf{r} and the columns of \mathbf{X} , so that $R = \cos(\mathbf{q})$ where \mathbf{q} is the angle between \mathbf{r} and the column space of \mathbf{X} . Thus $R = 0$ implies $\mathbf{q} = \pi/2 = 90^\circ$, and the corresponding \mathbf{r} contains none of the structure of \mathbf{X} . The formula defining R is: $R^2 = \mathbf{r} \mathbf{X} \mathbf{X}^+ \mathbf{r} / \mathbf{r} \mathbf{r}$ when $\mathbf{r} \neq \mathbf{0}$, and $R = 0$ otherwise. Thus it is clear that $R^0 = R(\mathbf{r}^0, \mathbf{X})$ is always zero, while $\hat{R} = R(\hat{\mathbf{r}}, \mathbf{X})$ is usually not zero because $\hat{\mathbf{r}}$ is orthogonal to the column space of $\mathbf{V}^{-1} \mathbf{X}$ rather than orthogonal to the column space of \mathbf{X} .

An unfortunate property of \hat{R} as a measure of pathology between \mathbf{X} and \mathbf{V} is that \hat{R} depends upon \mathbf{y} . Thus we now consider maximizing \hat{R} (minimizing $\hat{\mathbf{q}}$) by choice of \mathbf{y} so as to measure the maximum possible conflict between \mathbf{V} and \mathbf{X} . Using a well known lemma on extrema of quadratic forms

[10, p.62, (i)], it follows that $\max(\hat{R}^2)$ is equal to the largest eigen value of the nonnegative definite matrix $\mathbf{Q}^# \mathbf{X} \mathbf{X}^+ \mathbf{Q}^#$, where $\mathbf{Q}^#$ is the orthogonal projection orthogonal to the column space of $\mathbf{V}^{-1} \mathbf{X}$, namely $\mathbf{Q}^# = \mathbf{I} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{V}^{-2} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1}$.

A few numerical values of $\max(\hat{R}^2)$ and $\min(\hat{\mathbf{q}})$ are given in the tabulation below for the case where \mathbf{V} is first order autoregressive [$v_{ij} = \mathbf{a}^{|i-j|} / (1 - \mathbf{a}^2)$] with coefficient $\alpha = 0.5, 0.7$ or 0.9 , the \mathbf{X} design matrix includes a mean and linear trend ($p = 2$), and the sample size is $n = 5, 15$ or 25 . Gauss-Markov residuals will not have curious properties when α is negative or small and positive, but residuals can tend to have a linear trend in their values when α is close to one.

α	$n = 5$	$n = 15$	$n = 25$
+0.9	0.479/ 46°	0.659/ 36°	0.684/ 34°
+0.7	0.259/ 59°	0.357/ 53°	0.357/ 53°
+0.5	0.109/ 71°	0.152/ 67°	0.129/ 69°

An interesting comparison can also be made by restricting attention to the case $p = 1$ in which $\mathbf{X} = \mathbf{x}$ is a single column vector. Then, assuming $\mathbf{S} = \mathbf{s}^2 \mathbf{V}$, the efficiency of \mathbf{b}^0 relative to $\hat{\mathbf{b}}$ is $\text{Eff}(\mathbf{b}^0) = D(\hat{\mathbf{b}} | \mathbf{X}) / D(\mathbf{b}^0 | \mathbf{X}) = (\mathbf{x}' \mathbf{x})^2 / [(\mathbf{x}' \mathbf{V} \mathbf{x})(\mathbf{x}' \mathbf{V}^{-1} \mathbf{x})]$, [2, Sect. 10.2.2, page 567], while $\max(\hat{R}^2) = \mathbf{x}' \mathbf{Q}^# \mathbf{x} = 1 - (\mathbf{x}' \mathbf{V}^{-1} \mathbf{x})^2 / [(\mathbf{x}' \mathbf{x})(\mathbf{x}' \mathbf{V}^{-2} \mathbf{x})]$, where this maximum is achieved at $\mathbf{y} = c \mathbf{Q}^# \mathbf{x} + (\mathbf{I} - \mathbf{Q}^#) \mathbf{z}$ for $c \neq 0$ and any \mathbf{z} . With \mathbf{V} fixed, consider now choice of \mathbf{x} so as to minimize $\text{Eff}(\mathbf{b}^0)$ or maximize $\max(\hat{R}^2)$. In the following, the eigen value and eigen vector decomposition of \mathbf{V} will be written as $\mathbf{V} = \lambda_1 \mathbf{g}_1 \mathbf{g}_1' + \dots + \lambda_n \mathbf{g}_n \mathbf{g}_n'$ where $\lambda_1 \geq \dots \geq \lambda_n > 0$, $\mathbf{g}_i' \mathbf{g}_i = 1$, and $\mathbf{g}_i' \mathbf{g}_j = 0$ for $i \neq j$.

Theorem 3: If $p = 1$ and $\mathbf{S} = \mathbf{s}^2 \mathbf{V}$, then

$$(i) \min_{\mathbf{x} \neq 0} [\text{Eff}(\mathbf{b}^0)] = 1 - \max_{\mathbf{x} \neq 0} [\max_{\mathbf{y}} (\hat{R}^2)] = 4\lambda_1 \lambda_n / (\lambda_1 + \lambda_n)^2$$

- (ii) $\min_{\mathbf{x} \neq 0} [\text{Eff}(\mathbf{b}^0)]$, is achieved at $\mathbf{x} = c (\mathbf{g}_1 \pm \mathbf{g}_n)$ for $c \neq 0$, and
- (iii) $\max_{\mathbf{x} \neq 0} [\max_y (\hat{R}^2)] = (\lambda_1 - \lambda_n)^2 / (\lambda_1 + \lambda_n)^2$ is achieved at $\mathbf{x} = c (\lambda_1^{1/2} \mathbf{g}_1 \pm \lambda_n^{1/2} \mathbf{g}_n)$ for $c \neq 0$.

Proof: Applying the Kantorovich Inequality [10, p. 74], to $\text{Eff}(\mathbf{b}^0) = (\mathbf{x}\hat{\mathbf{C}}\mathbf{x})^2 / [(\mathbf{x}'\mathbf{V}\mathbf{x})(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})]$, yields $\text{Eff}(\mathbf{b}^0) \geq 4\lambda_1\lambda_n / (\lambda_1 + \lambda_n)^2$. Anderson [2, p. 569], uses a probabilistic form of the Kantorovich Inequality and comments upon the attainment, (ii), of this lower bound. An earlier reference for (ii) is Golub [5, p. 984]. Now note that $1 - \max (\hat{R}^2) = (\mathbf{z}\hat{\mathbf{C}}\mathbf{z})^2 / [(\mathbf{z}'\mathbf{V}\mathbf{z})(\mathbf{z}'\mathbf{V}^{-1}\mathbf{z})]$ for $\mathbf{z} = \mathbf{V}^{-1/2}\mathbf{x}$. Thus the remainder of (i) follows and, by (ii), $\max[\max(\hat{R}^2)]$ is attained at $\mathbf{z} = c (\mathbf{g}_1 \pm \mathbf{g}_n)$. The corresponding $\mathbf{x} = \mathbf{V}^{1/2}\mathbf{z}$ is that of (iii).

Theorem 3 shows that the smaller is the variance of $\hat{\mathbf{b}}$ relative to that of \mathbf{b}^0 , the larger can be the correlation between $\hat{\mathbf{r}}$ and \mathbf{x} . It is easily verified that $\text{Eff}(\mathbf{b}^0)$ at \mathbf{x} of (iii) equals one minus $\max(\hat{R}^2)$ at \mathbf{x} of (ii) equals $(\lambda_1 + \lambda_n)^2 / [2(\lambda_1^2 + \lambda_n^2)]$.

6. A TRANSFORMATIONAL INVARIANCE MOTIVATION FOR DIAGONALLY WEIGHTED LEAST SQUARES

A commonly seen motivation for Gauss-Markov estimation is the following. Let \mathbf{T} be a “square root” of \mathbf{V} in the sense that $\mathbf{V} = \mathbf{T}\mathbf{T}'$, and note that \mathbf{T} is not uniquely determined in the sense that $\mathbf{T}\mathbf{G}$ is another such square root for any orthogonal matrix \mathbf{G} . Now consider transforming both the \mathbf{y} data and the \mathbf{X} matrix so that $E(\mathbf{T}^{-1}\mathbf{y}|\mathbf{X}) = \mathbf{T}^{-1}\mathbf{X}\mathbf{b}$ and $D(\mathbf{T}^{-1}\mathbf{y}|\mathbf{X}) = \mathbf{s}^2\mathbf{I}$, which is the uncorrelated, homoscedastic observations case. It is then observed that the Gauss-Markov

estimator, $\hat{\mathbf{b}}$, in the original model is equivalent to the ordinary least squares estimator on the transformed data and design.

Note, however, that the original and transformed problems are equivalent only for estimation of \mathbf{b} and \mathbf{s}^2 . Since properties of residuals are being considered here, it is particularly disturbing that, unless \mathbf{V} and \mathbf{T} are diagonal matrices, the preceding transformational motivation for $\hat{\mathbf{b}}$ destroys the identity of the original observations, design points (rows of \mathbf{X}), and residuals by taking linear combinations of their original elements. The residuals corresponding to $\hat{\mathbf{b}}$ are, after the transformation, ordinary least squares residuals with optimal properties, but the transformation $\mathbf{r}^o = \mathbf{T}^{-1}\hat{\mathbf{r}}$ can be quite complicated when \mathbf{T} is not diagonal.

A referee points out that taking \mathbf{G} so that $\mathbf{T}\mathbf{G}$ is the positive definite square root of \mathbf{V} has the advantage that $D[(\mathbf{I} - \mathbf{G}\mathbf{T}^{-1})\mathbf{y} | \mathbf{X}]$ is then minimized in the sense of Theil [11] and Grossman and Styan [6]. In this weak sense, the elements of $\hat{\mathbf{r}}$ and of $\mathbf{r}^o = \mathbf{V}^{-1/2}\hat{\mathbf{r}}$ correspond and have the same “identity.” However, the general question of how to choose \mathbf{G} so that the rows of $\mathbf{G}'\mathbf{T}^{-1}\mathbf{y}$ and of $\mathbf{G}'\mathbf{T}^{-1}\mathbf{X}$ can be given simple, meaningful labels in terms of their original rows appears to be a formidable problem.

On the other hand, there is an obvious diagonal choice for \mathbf{T} that makes the elements $\mathbf{T}^{-1}\mathbf{y}$ homoscedastic (but not generally uncorrelated) and completely preserves the identity of the rows of \mathbf{y} and \mathbf{X} . This choice is $\mathbf{T}^2 = \text{Diag}(\mathbf{V})$, and the estimator of \mathbf{b} in the original model which is equivalent to ordinary least squares on the diagonally transformed problem is diagonally weighted least squares, $\mathbf{b}^D = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ where the “weight matrix” is $\mathbf{W} = [\text{Diag}(\mathbf{V})]^{-1}$.

An unfortunate property of OLS estimation is that this estimator is not invariant under simple rescalings of the rows of \mathbf{y} and \mathbf{X} ; these are the cases where $y_j \rightarrow c_j y_j$ and $x_{ij} \rightarrow c_j x_{ij}$ for $1 \leq j \leq p$, $1 \leq i \leq n$, and nonzero constants (c_1, \dots, c_p) . Note that diagonally weighted least squares estimation has the following optimality property: weighted least squares estimators using the diagonal weight matrix $\mathbf{W} = [\text{Diag}(\mathbf{V})]^{-1}$ are invariant under simple rescalings of the observations and are equivalent to ordinary least squares estimators on homoscedastic data so transformed as to preserve the labels on the observations.

The residuals corresponding to a positive definite diagonal choice for \mathbf{W} cannot display two curious properties which Gauss-Markov residuals ($\mathbf{W} = \mathbf{V}^{-1}$) can possess when \mathbf{V} is not diagonal. First, if \mathbf{W} is diagonal and \mathbf{X} contains the mean effect, $\mathbf{1}$, as one of its columns, all of the elements of \mathbf{r}^D cannot have the same numerical sign because $\mathbf{1}'\mathbf{W}^{-1}\mathbf{Q} = \mathbf{0}'$ implies that $\sum_{i=1}^n r_i^D / w_{ii} = 0$. On the other hand, Gauss-Markov residuals may all have the same numerical sign if all error variables are assumed to be highly positively correlated. Secondly, when fitting a linear trend, a diagonal choice of \mathbf{W} cannot produce all positive residuals on one side of some point and all negative residuals on the other side of that same point, which can happen with $\hat{\mathbf{r}}$ when the two groups of error variables were assumed to be highly negatively correlated.

REFERENCES

- [1] Aitken, A.C. "On Least Squares and Linear Combinations of Observations." *Proceedings of the Royal Society of Edinburgh*, Series A, 55 (1935), 42-8.
- [2] Anderson, T.W., **The Statistical Analysis of Time Series**, New York: John Wiley and Sons, Inc., 1971, 566-71.
- [3] Canner, P.L., "Some Curious Results Using Minimum Variance Linear Unbiased Estimators." *The American Statistician* 23, No. 5 (December 1969), 39-40.
- [4] Draper, N.R. and Smith, H., "The Examination of Residuals," **Applied Regression Analysis**, New York: John Wiley and Sons, Inc., 1966, 86-103.
- [5] Golub, G.H., "Comparisons of the Variance of Minimum Variance and Weighted Least Squares Regression Coefficients." *Annals of Mathematical Statistics* 34 (September 1963), 984-91.
- [6] Grossman, S.I. and Styan, G.P.H., "Optimality Properties of Theil's BLUS Residuals." *Journal of the American Statistical Association* 67 (September 1972), 672-3.
- [7] Hoerl, A.E. and Kennard, R.W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (February 1970), 55-67.
- [8] Larsen, W.A. and McCleary, S.J., "The Use of Partial Residual Plots in Regression Analysis." *Technometrics* 14 (August 1972), 781-90.
- [9] Marquardt, D.W., "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." *Technometrics* 12 (August 1970), 591-612.
- [10] Rao C.R., **Linear Statistical Inference and Its Applications**, Second Edition, New York: John Wiley and Sons, Inc., 1973.

[11] Theil, H., "The Analysis of Disturbances in Regression Analysis." *Journal of the American Statistical Association* 60 (December 1965), 1067-79.

[12] Wilk, M.B. and Gnanadesikan, R., "Probability Plotting Methods for the Analysis of Data." *Biometrika* 55 (March 1968), 1-17.

APPENDIX: DOUBLE DICHOTOMY OF RESULTS

Model Dichotomies	Expectation Model Correct: $\mu = X\beta$ for some β	Expectation Model Incorrect: $\mu \neq X\beta$ for any β
Dispersion Model Correct: $\Sigma = \sigma^2V$	Case One	Case Three
Dispersion Model Incorrect: $\Sigma \neq \sigma^2V$	Case Two	Case Four

- Only in case one is Gauss-Markov estimation, $\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y$, optimal (BLU.) This is also the only case where Gauss-Markov residuals, \hat{r} , have minimum mean squared error deviation from the true vector of errors.
- In all four cases, ordinary least squares (OLS) residuals are optimal estimates of lack-of-fit (Theorem 2.)
- In all four cases, diagonally weighted least squares provides (i) an estimate of β that is invariant under rescalings of observations and (ii) residuals that can be transformed to optimal OLS residuals without losing their individual identities.
- In cases one and two (where $\mu = X\beta$), the coefficient of variation of the residual vector corresponding to any linear, unbiased estimator of β (i.e. the corresponding L matrix is a generalized inverse of X) is $+\infty$ (Theorem 1.)
- In cases one and three (where $\Sigma = \sigma^2V$), $\text{Eff}(b^o) \geq 1 - \max[\max(\hat{R}^2)]$ when the rank of the X matrix is 1 (Theorem 3.)

At the time of publication (June 1975), the author (R. L. Obenchain) was a member of the technical staff, Applied Statistics Department, Bell Telephone Laboratories, HO-2E425, Holmdel, N.J. 07733. The author acknowledged he was indebted to C.L. Mallows for comments on previous descriptions of the article. He also thanked the referees for helpful suggestions, especially with respect to Theorems 1 and 3.