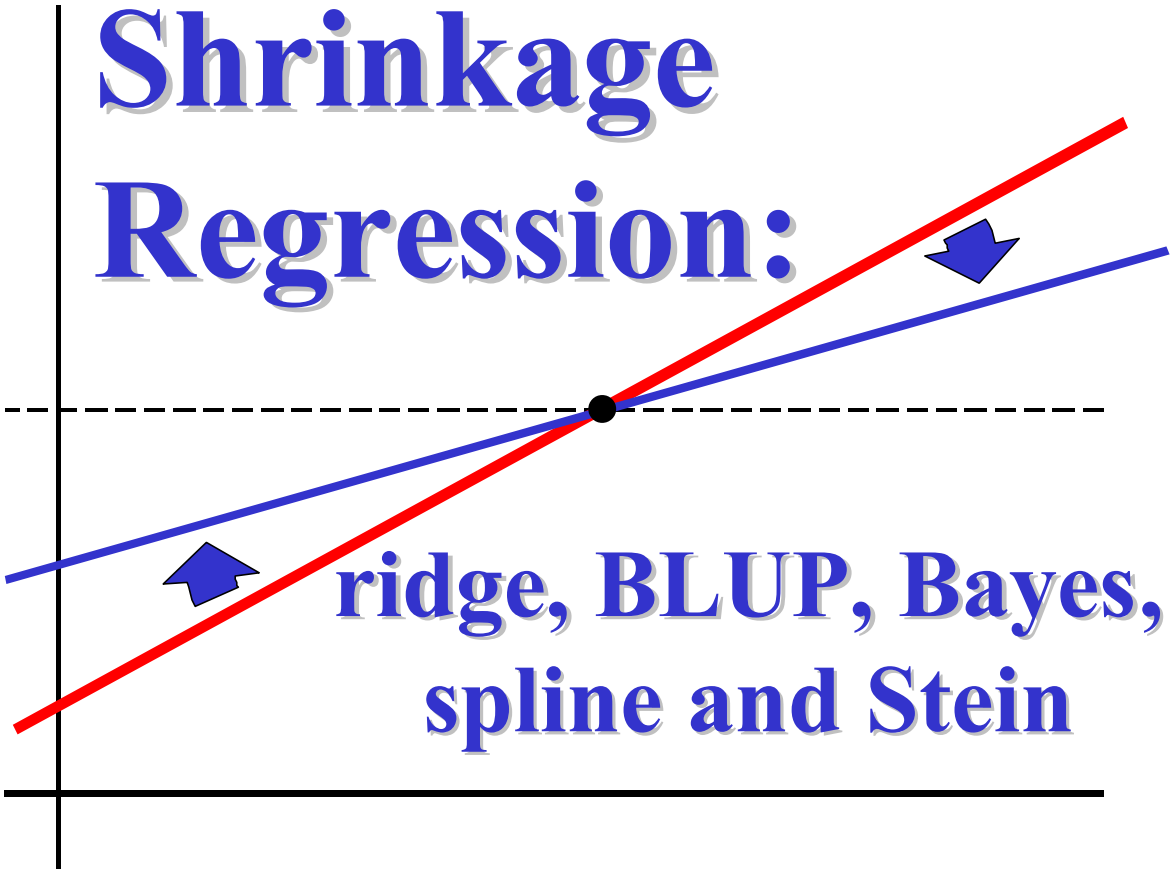


Chapter 02: Basic Linear Model Concepts

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 2: BASIC LINEAR MODEL CONCEPTS

Multiple linear regression models quantify the relationship of a “response" variable, Y, to P given non-constant “regressor" variables (also called “predictor" or “independent" variables), $X_1, \ldots, X_P$. Wherever reasonable in this book, we will use standardized symbols and vector-matrix notation. For example, the N observed response values are formed into an N by 1 column vector, **y**, and the regressor values are formed into an N by P matrix, **X**. In other words, rows represent observations, and columns represent variables. Our linear model then becomes . . .

Conditional Expectation of **y** given **X**: $\quad E(\mathbf{y} \mid \mathbf{X}) = \mathbf{1}\,\mu + \mathbf{X}\,\boldsymbol{\beta}$ $\quad \{ 2.1 \}$

Conditional Variance of **y** given **X**: $\quad V(\mathbf{y} \mid \mathbf{X}) = \sigma^2\,\mathbf{I}$ $\quad \{ 2.2 \}$

where

$\mathbf{1}$ = column vector of N ones,
μ = unknown intercept term (scalar valued),
$\boldsymbol{\beta}$ = column vector of P unknown, true regression coefficients, and
σ^2 = unknown residual variance (non-negative scalar).

Individual regression coefficients may be visualized as being either fixed or random. A multiple linear regression model with both fixed and random coefficients is said to be a mixed model.

2.1 Centered Variables

The above linear model can be restated in terms of centered variables as follows. Suppose that the mean response value has been subtracted from each row of the response vector, so that $\mathbf{1}^T\mathbf{y} = 0$, and the row vector of regressor (column) means, $\bar{\mathbf{x}}^T$, has been subtracted from each row of the regressor matrix, so that $\mathbf{1}^T\mathbf{X} = \mathbf{0}^T$. Note that centering is equivalent to replacing the **y** vector by $(\mathbf{I} - \mathbf{1}\,\mathbf{1}^T/N)\,\mathbf{y}$ and replacing the **X** matrix by $(\mathbf{I} - \mathbf{1}\,\mathbf{1}^T/N)\,\mathbf{X}$. Our pair of model equations then become . . .

Conditional Expectation of **y** [centered] given **X** [centered] :

$$E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\beta \qquad \{ 2.3 \}$$

Conditional Variance of **y** [centered] given **X** [centered] :

$$V(\mathbf{y} \mid \mathbf{X}) = \sigma^2 (\mathbf{I} - \mathbf{1}\mathbf{1}^T / N) \qquad \{ 2.4 \}$$

Notice the key difference between equations { 2.2 } and { 2.4 }. In equation { 2.2 }, the variance matrix, $\sigma^2 \mathbf{I}$, expresses the familiar condition that the response disturbance terms, $\mathbf{y} - E(\mathbf{y} \mid \mathbf{X})$, although possibly not statistically independent and identically distributed, are at least uncorrelated with a common variance (the so-called "homoscedastic" observations case.) The corresponding notation for the variance matrix of the centered response vector is $\sigma^2(\mathbf{I} - \mathbf{1}\mathbf{1}^T / N)$ of equation { 2.4 }, which will be much less familiar to many readers. This notation simply reminds us that the set of N deviations of response values from their common mean value, $\mathbf{y} - \bar{y}\mathbf{1}$, have the variance-covariance structure of "interchangable" (or "exchangeable") random variables. These deviations cannot be uncorrelated; the "centering" process gave them a non-random sum of ZERO.

The potential advantages of using centered-variable notation are illustrated by the following pair of formulas for the "ordinary least squares estimator," $\mathbf{b}^o$, of the P elements of β in the so-called "full rank" case. In un-centered notation,

$$\mathbf{b}^o = [\mathbf{X}^T (\mathbf{I} - \mathbf{1}\mathbf{1}^T / N) \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{1}\mathbf{1}^T / N) \mathbf{y} ; \qquad \{ 2.5 \}$$

but, when re-expressed in terms of centered variables,

$$\mathbf{b}^o = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} . \qquad \{ 2.6 \}$$

On the other hand, it is not really necessary to assume (as in formulas { 2.5 } and { 2.6 }) that the matrix of centered sums-of-squares and cross-products, $\mathbf{X}^T(\mathbf{I} - \mathbf{1}\mathbf{1}^T / N)\mathbf{X}$, is of full rank (P) in order to define $\mathbf{b}^o$; see equation { 2.9 } below. Our point here is simply that the second expression illustrates the great gain in notational simplicity that can result from adopting the convention that all variables have been centered. Therefore, all remaining formulas in these notes are displayed tacitly assuming, unless specifically stated otherwise, that all variables have been centered by subtracting off the appropriate column mean from every row of the corresponding response vector and/or regressor matrix.

2.2 The Special Case of Uncorrelated Regressors

In the special case where the regressor variables are uncorrelated, the centered $\mathbf{X}^T\mathbf{X}$ matrix will be diagonal. And the least squares solution vector of equation { 2.6 } then yields individual coefficients of a particularly simple form. Namely, the i-th coefficient is then

$$b_i^o \;=\; \frac{\mathbf{x}_i^T\,\mathbf{y}}{\mathbf{x}_i^T\,\mathbf{x}_i} \;=\; \sqrt{\frac{\mathbf{y}^T\,\mathbf{y}}{\mathbf{x}_i^T\,\mathbf{x}_i}}\;\; r_{yx_i}\,, \qquad \{\,2.7\,\}$$

where r_{yx_i} denotes the "marginal" correlation between the response vector, $\mathbf{y}$, and the i-th regressor variable, $\mathbf{x}_i$ = i-th column of the $\mathbf{X}$ matrix.

Models with uncorrelated-regressors rarely occur in actual practice (except, possibly, in "designed" experiments), but it is interesting to note how least squares estimates are constructed in this limiting case. Note that each least squares regression coefficient is directly proportional to the corresponding marginal correlation in { 2.4 }; in particular, a fitted coefficient is guaranteed to have the same numerical sign as the marginal correlation between regressor and response coordinates. On the other hand, note that the numerical magnitude of an individual coefficient can depend as much upon your choice of scaling, $\sqrt{\mathbf{x}_i^T\mathbf{x}_i}$, for its regressor coordinates as it does upon the marginal correlation of that regressor with the response. (The scaling of the response variable, embodied by the $\sqrt{\mathbf{y}^T\mathbf{y}}$ term, applies equally to all coefficients.) Specifically, the i-th coefficient can be relatively large simply because the observed coordinates along the i-th regressor axis have relatively small "spread" (a small $\sqrt{\mathbf{x}_i^T\mathbf{x}_i}$ term) rather than because the corresponding marginal correlation is relatively large. As we will see later (in equation { 2.23 }), the statistical significance of a fitted regression coefficient depends only upon the corresponding correlation coefficient, but its numerical size can be "distorted" by any unusual, extreme choice for the scaling of its regressor/predictor coordinates.

2.3 Canonical Form of Regressors

Here, we illustrate how the basic correlation-spread relationships observed in { 2.7 } apply to the general case where regressors are usually intercorrelated. However, to see these relationships clearly in equation { 2.16 }, we will first need to "rotate" axes to canonical form. Our notation will make frequent reference to the component parts, H, Λ and G, of the singular value decomposition of the centered regressor matrix:

$$\mathbf{X} \;=\; (\,\mathbf{I} \;-\; \mathbf{1}\,\mathbf{1}^T/\,N\,)\,\mathbf{X} \;=\; \mathbf{H}\,\mathbf{\Lambda}^{1/2}\,\mathbf{G}^T, \qquad \{\,2.8\,\}$$

where

H is a semi-orthogonal (N by R) matrix of standardized "principal coordinates of **X**,"

G is a semi-orthogonal (P by R) matrix of "principal axis direction cosines for **X**,"

$\mathbf{\Lambda}^{1/2}$ is a diagonal (R by R) matrix of "ordered singular values of **X**," and

R denotes the RANK of **X**, where $1 \leq R \leq P \leq N$.

Like the centered **X** matrix, the principal coordinates matrix, **H**, has a zero mean vector, $\mathbf{1}^T\mathbf{H} = \mathbf{0}^T$. But the **H** matrix also has the properties that $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ (R by R) and $\mathbf{H}\mathbf{H}^T$ is the orthogonal projection matrix onto the column space of **X**, which is a uniquely determined, symmetric, and idempotent matrix that is N by N of rank R, Rao(1973), page 47.

The principal axis direction cosine matrix, **G**, always has columns ($\vec{\mathbf{g}}_1, \ldots, \vec{\mathbf{g}}_R$) that are mutually orthogonal and of length one: $\mathbf{G}^T\mathbf{G} = \mathbf{I}$ (R by R.) In the so-called "full rank" case ($R = P$), $\mathbf{G}\mathbf{G}^T$ is also a P by P identity matrix (i.e., **G** is orthogonal rather than simply semi-orthogonal.) But, when R is strictly less than P, $\mathbf{G}\mathbf{G}^T$ is the orthogonal projection matrix for the row space of **X**, which is a uniquely determined, symmetric, and idempotent matrix of rank R that is $P \times P$.

The ordered singular values of **X** (from upper-left to lower-right along the main diagonal of $\mathbf{\Lambda}^{1/2}$) are

$$\lambda_1^{1/2} \geq \lambda_2^{1/2} \geq \ldots \lambda_R^{1/2} > 0. \qquad \{ 2.9 \}$$

Note, specifically, that each of the first R singular values is strictly positive (greater than zero.)

When **X** is NOT of full (column) rank (i.e., $R < P$), the final $P - R$ singular values of **X** are exact zeros, which were dropped from equations { 2.8 } and { 2.9 }. When $P - R \geq 2$, there would be no unique way to add 2 or more columns to the **H** and **G** matrices. Similarly, when two (or more) of the ordered singular values in { 2.9 } are exactly equal, the corresponding columns of **H** and **G** are also not uniquely determined. (It is always possible to, say, multiply all of the elements in any column of **H** by -1 if one also multiplies all elements in the corresponding column of **G** by -1. This sort of modification of the singular valve decomposition is too "trivial" to cause any real concern about "uniqueness.")

Non-uniqueness caused by one or more zero singular values carries over to estimates of the $\boldsymbol{\beta}$ coefficient vector, at least when estimates are viewed as "solutions" to under-determined "normal equations," $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$. Twenty years ago, $\boldsymbol{\beta}$ was commonly said to be "not estimable" whenever $R < P$. Today, it is apparently more common to adopt the convention that "the" least squares estimator of $\boldsymbol{\beta}$ is always (even in the less-than-full-rank-case) defined to be...

$$\mathbf{b}^o \equiv \mathbf{X}^+\mathbf{y}\,, \qquad \{ 2.10 \}$$

where the $+$ superscript denotes the (unique) Moore-Penrose inverse of the (centered) regressor matrix, **X** , Rao(1973), page 26. It is straightforward to show, using { 2.8 } and { 2.9

}, that $\mathbf{X}^{+} = \mathbf{G}\,\mathbf{\Lambda}^{-1/2}\,\mathbf{H}^{\mathrm{T}}$ when R is either less than or equal to P. Other implications of this factorization are explored in great detail below.

NUMERICAL EXAMPLE:

Table 2.1 below lists the data for a small numerical example with $N = 10$ observations on $P = 2$ regressor variables and a response variable.

Table 2.1 A Small Numerical Example

X1	X2	Y
− 1.67	− 1.68	− 1.58
− 1.11	− 0.34	− 1.06
− 0.58	− 1.35	− 0.53
− 0.28	− 0.21	− 0.79
− 0.54	0.00	− 0.48
0.28	− 0.34	0.74
0.56	0.99	1.06
0.84	0.67	0.79
1.11	1.35	0.53
1.39	0.91	1.32

Besides being centered as explained in §2.1, so that the mean value in each column is 0, the numerical values of each variable have also been placed on a standardized scale. This rescaling involves dividing each original predictor coordinate by the sample standard deviation of that variable. As a result, the sum-of-squares of the $N = 10$ values in each column of Table 2.1 is $N - 1 = 9$. And the pairwise sample correlations between variables are

	X_1	X_2	Y
X_1	1.0000		
X_2	0.8683	1.0000	
Y	0.9401	0.7901	1.0000 .

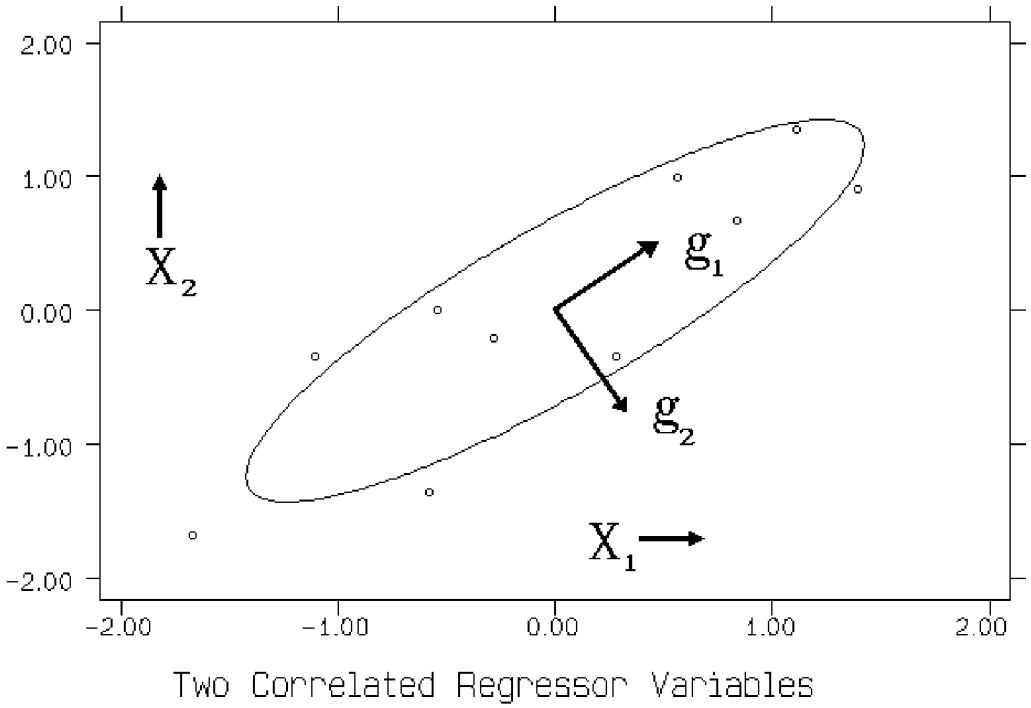
Because regressor variables have exactly equal variance following this rescaling, principal axes for the bivariate ($P = R = 2$) case will always be rotated to exactly a 45° angle relative to the given regressor axes. Although not uniquely determined, each element of the 2×2 matrix, $\mathbf{G}$, of principal axis direction cosines will be $\pm \sqrt{2}$. For example, one possible choices is

$$G = \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix}.$$

By the way, rescaling regressors to have equal variance when $P \geq 3$ does not necessarily yield a **G** matrix of any special form (like it does here in the bivariate case, $P = 2$.)

Figure 2.1 displays a scatter-plot of the X_1 and X_2 regressor coordinates from Table 2.1 and the $\vec{g}_1$ and $\vec{g}_2$ axes corresponding to the above choice for the columns of **G** . Furthermore, a constant-density ellipse of a hypothetical bivariate-normal distribution for a pair of standardized variables with the same inter-correlation, $\widehat{\rho}_{12} = +0.8683$, as that observed between X_1 and X_2 is also shown in Figure 2.1.

Figure 2.1 Given Regressor Coordinates and Principal Axes



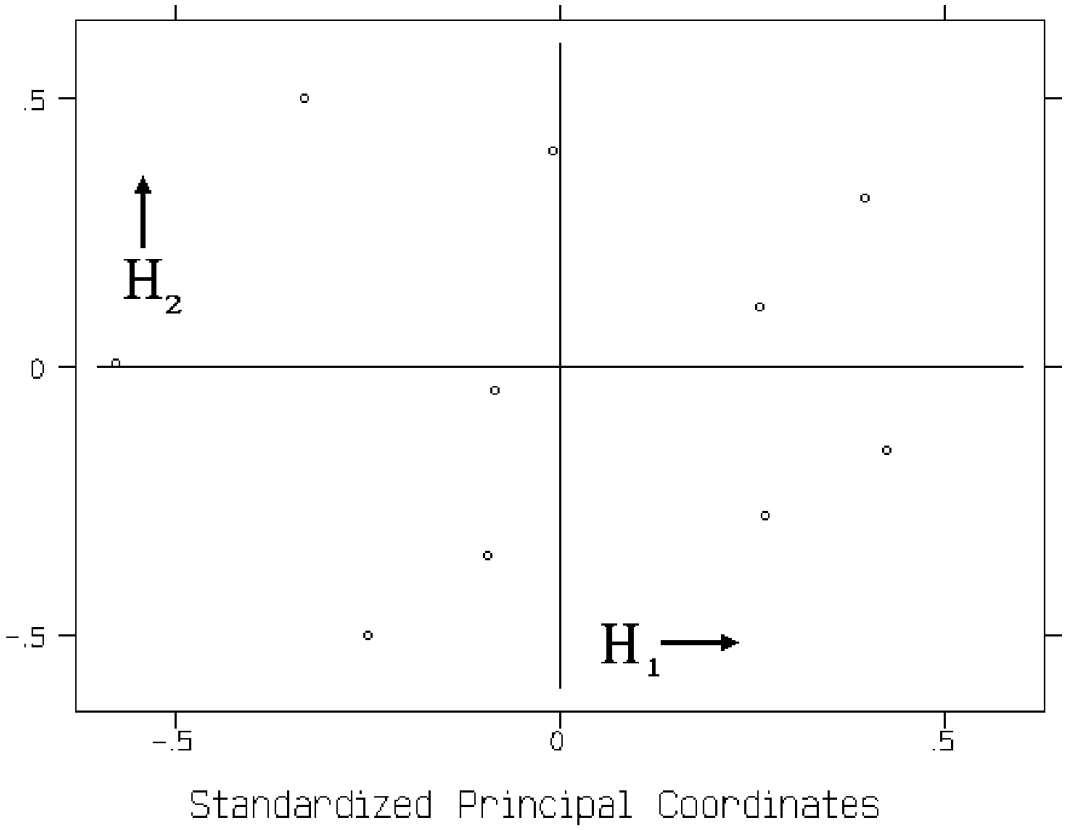
The singular values of the **X** matrix are $\lambda_1^{1/2} = \sqrt{(N-1)\cdot(1+\widehat{\rho}_{12})} = 4.1006$ and $\lambda_2^{1/2} = \sqrt{(N-1)\cdot(1-\widehat{\rho}_{12})} = 1.0886$. And the 10×2 **H** matrix of standardized regressor principal coordinates is

$$\mathbf{H} = \begin{matrix} -0.5778 & 0.0059 \\ -0.2502 & -0.5006 \\ -0.3329 & 0.4999 \\ -0.0845 & -0.0456 \\ -0.0932 & -0.3510 \\ -0.0103 & 0.4028 \\ 0.2673 & -0.2791 \\ 0.2605 & 0.1107 \end{matrix}$$

0.4243 – 0.1555
0.3968 0.3123

Figure 2.2 displays a scatter-plot of these standardized H_1 and H_2 coordinates, showing that they are uncorrelated with equal variance. (The sum-of-squares of the N = 10 values in each column of the **H** matrix is 1.)

Figure 2.2 Regressor Principal Coordinates



2.4 Numerical versus Statistical Ill-Conditioning

Terminology: A multiple regression problem is called either EXACTLY SINGULAR or NUMERICALLY ILL-CONDITIONED whenever the matrix of centered regressor coordinates is less than full rank.

In actual regression practice, exact singularities and/or ties among singular values are rather rare (except, again, in designed “orthogonal" experiments.) Anyway, regression practitioners usually find themselves in the common situation where the singular values of X are distinct...

$$\lambda_1^{1/2} > \lambda_2^{1/2} > \ldots > \lambda_P^{1/2} > 0 . \qquad \{ 2.11 \}$$

All of the component parts [$\mathbf{H}$, $\mathbf{\Lambda}$ and $\mathbf{G}$] of the singular value decomposition are “essentially" uniquely determined when { 2.11 } holds.

Statistical ill-conditioning is a much more common problem for regression practitioners than is numerical ill-conditioning. In fact, statistical ill-conditioning is almost always present in at least some very weak form whenever data collection had failed to follow a well-planned design. We can give the following “qualitative" definition now; much greater insight into this general topic will be provided below in Section §2.6.

A multiple regression problem is said to be STATISTICALLY ILL-CONDITIONED when the trailing singular values, $\lambda_P^{1/2}$, $\lambda_{P-1}^{1/2}, \ldots$, of the centered regressor matrix, $\mathbf{X}$, are numerically small compared to its leading singular values, $\lambda_1^{1/2}$, $\lambda_2^{1/2}, \ldots$

2.5 Eigen Decompositions

The familiar eigenvalue-eigenvector decomposition of the regressor adjusted sums-of-squares and cross-products matrix, $\mathbf{X}^T \mathbf{X}$, is closely related to the singular value decomposition of equation { 2.7 } in the sense that:

$$\mathbf{X}^T\mathbf{X} = \mathbf{X}^T(\mathbf{I} - \mathbf{1}\mathbf{1}^T/ N)\mathbf{X} = \mathbf{G}\mathbf{\Lambda}\mathbf{G}^T. \qquad \{ 2.12 \}$$

This well-known decomposition technique is useful in numerical computations where N is much, much larger than P. Rather than attack the N by P centered regressor matrix, $\mathbf{X}$, via the singular value decomposition when N >> P, one can use the eigenvalue-eigenvector approach on a much smaller (P by P) matrix to calculate only $\mathbf{\Lambda}$ and $\mathbf{G}$. Later, when actually needed, principal coordinates can always be computed (if only approximately) by simple matrix multiplication, $\mathbf{H} = \mathbf{X}\,\mathbf{G}\,\mathbf{\Lambda}^{-1/2}$.

The ordered eigenvalues of $\mathbf{X}^T \mathbf{X}$, which are denoted by

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_R > 0, \qquad \{ 2.13 \}$$

are the "adjusted" (via centering) sums-of-squares of **X** coordinates along their principal axes. In other words, an eigenvalue is (N – 1) times the sample variance along a certain direction in P-dimensional regressor space. Similarly, the corresponding singular value (square root of the eigenvalue) is $\sqrt{N-1}$ times a sample standard deviation along that same direction in regressor space.

2.6 The Uncorrelated Components of Least Squares

The notation and terminology introduced above allows us to examine, in great detail, the basic structure of the least squares estimator of $\boldsymbol{\beta}$. Specifically, we can rewrite equation { 2.9 } as

$$\mathbf{b}^o = \mathbf{G}\,\mathbf{c}\,, \qquad \{ 2.14 \}$$

where **G** is the semi-orthogonal direction cosines matrix for the principal axes of the centered regressors and **c** is the R × 1 column vector containing the uncorrelated components of $\mathbf{b}^o$. Thus, by definition, the vector of uncorrelated components is of the form:

$$\begin{aligned} \mathbf{c} &\equiv \mathbf{G}^T \mathbf{b}^o, \\ &= \boldsymbol{\Lambda}^{-1/2}\,\mathbf{H}^T\,\mathbf{y}\,, && \{ 2.15 \} \\ &= \sqrt{\mathbf{y}^T\mathbf{y}}\cdot\boldsymbol{\Lambda}^{-1/2}\,\mathbf{r}\,. && \{ 2.16 \} \end{aligned}$$

In equation { 2.16 } , $\mathbf{r} = (r_{yi})$ represents the column vector of principal correlations between the response vector, **y**, and the columns of the principal axis regressor coordinate matrix, **H**. Specifically, $\mathbf{r} = \mathbf{H}^T\,\mathbf{y} / \sqrt{\mathbf{y}^T\mathbf{y}}$, which is an R × 1 vector. As is stressed below, equation { 2.16 } has a lot to say about the basic nature of statistical ill-conditioning in multiple linear regression models!

However, we first observe that $\mathbf{b}^o$ will be an unbiased estimator of $\boldsymbol{\beta}$ when rank(**X**) = R = P as long as the expectation models, { 2.1 } and { 2.3 }, are "correct" statistical models for the data at hand. In other words, whenever β is estimable we have

$$E (\mathbf{b}^o \mid X) = \mathbf{X}^+\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}\,. \qquad \{ 2.17 \}$$

Furthermore, if the variance models, { 2.2 } and { 2.4 }, are "correct" models, the variance matrix of $\mathbf{b}^o$ will be of the form:

$$V (\mathbf{b}^o \mid \mathbf{X}) = \sigma^2\,\mathbf{X}^+\,\mathbf{X}^{+T} = \sigma^2\,(\mathbf{X}^T\,\mathbf{X})^+. \qquad \{ 2.18 \}$$

Of course, equation { 2.18 } becomes simply $V(\mathbf{b}^o \mid \mathbf{X}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ when $R = \text{rank}(\mathbf{X}) = P$.

The corresponding expectation vector and variance matrix for the uncorrelated components are

$$E(\mathbf{c} \mid \mathbf{X}) \equiv \gamma = \mathbf{G}^T \beta , \qquad \{ 2.19 \}$$

and

$$V(\mathbf{c} \mid \mathbf{X}) = \sigma^2 \mathbf{\Lambda}^{-1} . \text{ (R by R)} \qquad \{ 2.20 \}$$

Note, specifically, that our terminology here is motivated by the observation that the elements of **c** are always uncorrelated (i.e. their variance matrix is always diagonal.)

Developing an appreciation for the implications of formula { 2.16 } is fundamental to understanding ill-conditioning as a dual numerical/statistical phenomenon. Therefore, let us now examine the individual elements of the uncorrelated component **c** vector. The i-th such element is...

$$c_i = \sqrt{\mathbf{y}^T \mathbf{y}} \cdot r_{yi} / \lambda_i^{1/2} . \qquad \{ 2.21 \}$$

Individual uncorrelated components can be relatively large, <u>numerically</u>, either because their corresponding principal correlation is relatively large or simply because their corresponding regressor singular value is relatively small.

On the other hand, note from equation { 2.20 } that the standard error of the i-th uncorrelated component, c_i , is σ divided by $\lambda_i^{1/2}$. In other words, the "relative" standard error of c_i is the known quantity $\lambda_i^{-1/2}$. Therefore, one's uncertainty about the size of each true component of β is inversely proportional to the spread in regressor coordinates along the corresponding principal axis.

NUMERICAL EXAMPLE: Let us now continue the numerical example ($P = 2$, $N = 10$) we introduced in section §2.3. The marginal correlations between the response **y** values and the two given regressor **X** columns were previously stated to be $\hat{\rho}_{y1} = +0.9401$ and $\hat{\rho}_{y2} = +0.7901$. The uncorrelated components of $\mathbf{b}^o$ are thus $c_1 = +0.654828$ and $c_2 = +0.805537$ in { 2.15 }, and the principal correlations of the response **y** values with the two sets of regressor principal **H** coordinates are $r_{y1} = +0.8951$ and $r_{y2} = +0.2923$ in { 2.16 }. Note that r_{y1} is more than 3 times larger than r_{y2} . Yet c_1 is smaller, numerically, than is c_2 . And we know why! The second component, c_2 , is greatly magnified because the singular value in its denominator, $\lambda_2^{1/2} = 1.089$, is much smaller than the singular value, $\lambda_1^{1/2} = 4.101$, in the denominator of c_1 .

2.7 Statistical Significance of Uncorrelated Components

The F-ratio for testing the hypothesis that $\gamma_i = 0$ (i.e. testing the "statistical significance" of the i-th estimated component, c_i) is of the form:

$$F_i = \frac{c_i^2 \, \lambda_i}{s^2} . \qquad \{ 2.22 \}$$

where the least-squares residual-mean-square, $s^2 = \mathbf{y}^T (\mathbf{I} - \mathbf{H} \mathbf{H}^T) \mathbf{y} / (N - R - 1)$, is the estimator of the error variance, σ^2.

But, wait a second! A little bit of algebra reveals that the F-statistic of { 2.22 } does not actually depend upon the λ_i or any of the other regressor eigenvalues or singular values! Using expression { 2.22 } for c_i, an expression equivalent to { 2.22 } is:

$$F_i = \frac{(N-R-1) \cdot r_{yi}^2}{(1 - R^2)} \qquad \{ 2.23 \}$$

where R is the rank of $\mathbf{X}$ and R^2 is the familiar R-squared statistic, which can be expressed as the following sum-of-squares of principal correlations:

$$R^2 = r_{y1}^2 + r_{y2}^2 + \ldots + r_{yR}^2 .$$

The t-statistic corresponding to { 2.23 } is

$$t_i = r_{yi} \cdot \sqrt{\frac{(N-R-1)}{(1 - R^2)}} . \qquad \{ 2.24 \}$$

Equations { 2.23 } and { 2.24 } remind us that . . .

Individual uncorrelated components CANNOT be judged relatively important, statistically, simply because they are large numerically. A component can be large simply because its corresponding regressor singular value is relatively small. Statistical significance depends only upon the regressor principal correlations.

NUMERICAL EXAMPLE:

Again, continuing the numerical example (P = 2, N = 10) from sections §2.3 and §2.6 , the R-squared statistic is $R^2 = 0.8951^2 + 0.2923^2 = 0.8866$. And the t-statistics for the individual uncorrelated components, 7.03 and 2.30 , are directly proportional to the corresponding principal correlations, 0.8951 and 0.2923 .

2.8 Predictions, Residuals and Linear Reparameterizations that remove Ill-Conditioning.

Like the above F-ratios and t-statistics, it is easily shown that least-squares predicted responses and residuals also do not depend upon theregressor eigenvalues or singular values, $\mathbf{\Lambda}$ of { 2.8 }. In fact, these quantities also do not depend upon the direction cosines, **G** of { 2.8 }. Specifically, the vector of [centered] least-squares predictions is

$$\mathbf{y}^o = \mathbf{X}\,\mathbf{b}^o = \mathbf{X}\,\mathbf{X}^+\mathbf{y} = \mathbf{H}\mathbf{\Lambda}^{1/2}\mathbf{G}^T\mathbf{G}\mathbf{\Lambda}^{-1/2}\mathbf{H}^T\mathbf{y} = \mathbf{H}\,\mathbf{H}^T\mathbf{y}\,,$$

where $\mathbf{H}\,\mathbf{H}^T$ is again the uniquely determined orthogonal projection matrix for the R-dimensional column space of the centered **X** matrix. The corresponding vector of [centered] residuals is

$$\mathbf{y} - \mathbf{y}^o = (\mathbf{I} - \mathbf{H}\,\mathbf{H}^T)\mathbf{y} = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/N - \mathbf{H}\,\mathbf{H}^T)\mathbf{y}\,.$$

These residuals are used to define s^2, the estimate of σ^2 used in equation { 2.22 }. Note that $(\mathbf{I} - \mathbf{1}\mathbf{1}^T/N - \mathbf{H}\,\mathbf{H}^T)$ is the uniquely determined orthogonal projection matrix for the (N – R – 1) dimensional linear subspace orthogonal to both **1** and the column space of the centered **X** matrix.

Although it is true that $\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{\Lambda}^{-1/2}$, this relationship does NOT establish that the **H** matrix of standardized principal coordinates actually depends upon either **G** or $\mathbf{\Lambda}$ (at least in the usual situation where all of the singular values of **X** are distinct.) Rather, our point-of-view is that the singular value decomposition of **X** in equation { 2.8 } simultaneously defines the **H, G** and $\mathbf{\Lambda}$ matrices. And the above arguments describe senses in which the **G** and $\mathbf{\Lambda}$ matrices can be "disregarded," leaving only the standardized **H** coordinates to define least squares predictions for both the response vector and the residual vector

In view of the above observations, let us now consider the following linear reparameterization of our original, ill-conditioned regression model, { 2.3 } and { 2.4 }. Specifically, the centered **X** matrix is now replaced by its semi-orthogonal (N by R) matrix of principal coordinates, **H** = **XA,** where the linear transformation matrix is $\mathbf{A} = \mathbf{G}\mathbf{\Lambda}^{-1/2}$.

Conditional Expectation of **y** [centered] given **H** [centered] :

$$E(\mathbf{y} \mid \mathbf{H}) = \mathbf{H}\boldsymbol{\alpha} \qquad \{ 2.25a \}$$

Conditional Variance of **y** [centered] given **H** [centered] :

$$V(\mathbf{y} \mid \mathbf{H}) = \sigma^2 (\mathbf{I} - \mathbf{1}\mathbf{1}^T / N) \qquad \{ 2.26a \}$$

We now argue that this reparameterized regression model displays absolutely no ill-conditioning (numerical or statistical) in the senses discussed above. First, note that $\boldsymbol{\alpha}$ is $R \times 1$ like $\boldsymbol{\gamma}$ rather than $P \times 1$ like $\boldsymbol{\beta}$. Furthermore, the implied **G** and $\boldsymbol{\Lambda}$ matrices for this reparameterized regression can both be taken to be $R \times R$ identity matrices. And the least squares estimate of $\boldsymbol{\alpha}$ in { 2.3a } is $\mathbf{a}^o = \mathbf{H}^+\mathbf{y} = \mathbf{H}^T\mathbf{y}$. Thus the i-th element of $\mathbf{a}^o$ is of the form

$$a_i^o = \sqrt{\mathbf{y}^T\mathbf{y}} \cdot r_{yi} = c_i \lambda_i^{1/2}, \qquad \{ 2.27a \}$$

and it follows that the elements of $\mathbf{a}^o$ are uncorrelated and homoscedastic (i.e. have a common variance of σ^2.)

Although completely free of ill-conditioning in the above senses, estimates for this reparameterized model are, none the less, closely related to corresponding estimates for the original, ill-conditioned model. Equation { 2.21a } shows that least squares estimates for regression coefficients differ only by a known scale factor. Furthermore, both models yield the exact same response predictions and residuals! Finally, the F-ratio or t-statistic for testing the significance of an individual $\boldsymbol{\alpha}$ coefficient estimate is identical to the F-ratio or t-statistic for testing the significance of the corresponding element of the least squares estimate of $\boldsymbol{\gamma}$ in the original model.

Ok, so what are some of the potential implications of the above observations?

2.8.1 Almost Irrelevant Information

My personal opinion is that the above observations show that estimation methodology designed to treat ill-conditioning in multiple regression models should have little or no effect on predicted responses or fitted residuals. Least squares is adequate (perhaps, even ideal) for these estimation tasks. The fact that the given regression parameterization is ill-conditioned is almost irrelevant when attention is restricted to only response predictions and residuals.

Methods designed to treat ill-conditioning should focus almost exclusively on problems associated with the high intercorrelations between least squares estimates of the elements of the $\boldsymbol{\beta}$ vector and/or the corresponding wildly heteroscedastic estimates for the uncorrelated components, $\boldsymbol{\gamma}$. After all, it is the given ill-conditioned **X** matrix that contains the regressor variables of genuine interest to you. You wanted to estimate their $\boldsymbol{\beta}$ coefficients in the first place because these are the variables available to you that may determine expected response, $E(\mathbf{y} \mid \mathbf{X})$. Your bad luck is that these **X** variables are highly intercorrelated in the only available

data. Furthermore, your interest in the true α coefficients corresponding to the $\mathbf{H} = \mathbf{XG\Lambda}^{-1/2}$ reparameterization is limited because this linear transformation of regressor variables is sufficiently complicated that you cannot easily visualize what it "means."

2.8.2 The Inverse Linear Transformation Restriction

Smith and Campbell(1980) argued, among other things, that the presence or absence of ill-conditioning in a multiple regression model should essentially be ignored. Specifically, they considered pairs of regressor variable reparameterizations, $\mathbf{X_1}$ and $\mathbf{X_2}$, that differ only due to an invertible linear transformation, $\mathbf{A}$, of the form $\mathbf{X_1 A = X_2}$. Elementary matrix manipulations then establish that $\mathbf{X_1}\boldsymbol{\beta_1} = \mathbf{X_1 A A^{-1}}\boldsymbol{\beta_1} = \mathbf{X_2}\boldsymbol{\beta_2}$. In other words, when a linear regression model $E(\mathbf{y}|\mathbf{X_1}) = \mathbf{X_1}\boldsymbol{\beta_1}$ is reparameterized to $E(\mathbf{y}|\mathbf{X}_2) = \mathbf{X}_2\boldsymbol{\beta}_2$, it is clear that the transformed regression coefficients must necessarily satisfy the "inverse transformation restriction" that $\boldsymbol{\beta}_2 = \mathbf{A}^{-1}\boldsymbol{\beta_1}$.

Because true regression coefficients always follow this restriction upon linear reparameterization, Smith and Campbell(1980) took the argument one step further by reasoning that statistical methodology "should" provide coefficient estimates that also satisfy this restriction. Least squares estimates do indeed always satisfy $\mathbf{b}_2^o = \mathbf{A}^{-1}\mathbf{b}_1^o$ following invertible linear regressor transformations of the form $\mathbf{X_1 A = X_2}$. Equation { 2.21a } illustrates this for the special case where the $\mathbf{A}$ reparameterization matrix is $\mathbf{G\Lambda}^{-1/2}$ and $\mathbf{A}^{-1} = \mathbf{\Lambda}^{+1/2}\mathbf{G}^{\mathrm{T}}$.

As we shall see in Chapter 3 on shrinkage regression fundamentals, "optimally" shrunken estimates generally do NOT satisfy this reparameterization restriction. This is the case because there is a non-linear relationship between the form and extent of ill-conditioning in a given parameterization and the extent of shrinkage that minimizes risk (expected or mean squared error loss) via explicit variance-bias trade-offs.

2.9 Signal-to-Noise Ratios

Perhaps there is a final “irony" here. The unknown, true non-centrality of the i-th variance ratio, F_i, is

$$\phi_i^2 \;=\; \gamma_i^2\,\lambda_i \,/\, \sigma^2 \;=\; \alpha_i^2 / \,\sigma^2. \qquad \{\ 2.28\ \}$$

Now ϕ_i^2 is a signal-to-noise ratio that plays a pivotal role in the theory of mean-squared-error optimal shrinkage along the i-th principal regressor axis. The statistical “power" parameters that control detection sensitivity for the true components of $\boldsymbol{\beta}$ are directly proportional to the spreads, the λ's, in regressor coordinates along their principal axes. Minimal spread therefore implies minimal power.

Deliberately DESIGNING a statistically ill-conditioned EXPERIMENT is almost surely unwise. These notes will concentrate upon methods useful in OBSERVATIONAL STUDIES where the regression practitioner has no hope of "controlling" or "guiding" the data collection process.

2.9 The Statistical Distribution of Principal Correlations

Even under "normal distribution theory," the exact statistical distribution of the principal correlations, $r_{y1}, \ldots, r_{yP}$, is usually not that of classical correlation coefficients. This is the case because the distribution theory of interest to us will almost always be conditional on the observed regressor values. From this point-of-view, the vector of principal correlations, $\mathbf{r} = \mathbf{H}^T \mathbf{y} / \sqrt{\mathbf{y}^T \mathbf{y}}$, is simply a known linear transformation (defined by the $\mathbf{H}^T$ matrix) of the version of the response vector, $\mathbf{y}$, that has been "rescaled" to be of unit length. The linear ($\mathbf{H}^T$) part of this transformation reduces dimensionality from N-dimensional response space down to R-dimensions. Therefore, if response values start out having a joint (multivariate) normal distribution given $\mathbf{X}$, the principal correlations end up having a "rescaled" normal distribution confined to an R-dimensional hypersphere, $\mathbf{r}^T \mathbf{r} \leq 1$, of unit radius. The principal correlations have equal variances, namely

$$V(r_{yi}) = \frac{(1-R^2)}{(N-R-1)} = \frac{s^2}{y^T y} \qquad \{ 2.29 \}$$

In our ($P = 2$, $N = 10$) numerical example, the two principal correlations each have standard error $\sqrt{(1 - 0.8866)/7} = 0.127291$.

2.10 When "Should" Coefficients have "Wrong" Signs?

Cases where a fitted least-squares regression coefficient has the "wrong" (unbelievable) numerical sign seem to arise most frequently in applications with "many" or, at least, "several" predictor (X) variables, as in the four-predictor gasoline-mileage example of Section §1.3. But this phenomenon can also be illustrated when there are only two predictors. Here in Section §2.10, we will explore only this most simple case ($P = R = 2$) of multiple regression. We start out with the usual theoretical model under which all three variables (Y, X_1 and X_2) have a joint, stochastic distribution. But we will end up relating our observations back to the common applications (of conditional inference) in which predictor coordinates are viewed as given.

NUMERICAL EXAMPLE:

Again, continuing the numerical example ($P = 2,\ N = 10$) we started in section §2.3 , the ordinary least-squares regression coefficients estimates are 1.03263 and -0.106568 . Thus the second coefficient does have the "wrong" sign in the sense we will be considering here. Because the two coefficients have equal variances in this case, the corresponding t-statistics are 4.02363 and -0.415237; thus, the second coefficient is not significantly different from zero.

2.10.1 Stochastic Response and Predictor Variables

Suppose that the joint distribution of Y, X_1 and X_2 has the vector of expected values

$$E \begin{bmatrix} Y \\ X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \mu_y \\ \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \{ 2.30 \}$$

and the matrix of variances

$$V \begin{bmatrix} Y \\ X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \sigma_y^2 & \rho_{y1}\sigma_y\sigma_1 & \rho_{y2}\sigma_y\sigma_2 \\ \rho_{y1}\sigma_y\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{y2}\sigma_y\sigma_2 & \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \qquad \{ 2.31 \}$$

where each μ term represents the mean value of a variable; each σ term represents the standard deviation (square root of the variance) of a variable; each ρ term represents a correlation between two variables; and the subscripts y, 1, and 2 correspond to the variables Y, X_1, and X_2 , respectively.

When a two-predictor regression model is to be fit to data, the elements of the E vector and the V matrix shown in equations { 2.27 } and { 2.28 } are simply replaced by their natural estimates, $\widehat{\mu}_j$, $\widehat{\sigma}_j$, and $\widehat{\rho}_{jk}$ for $j = y, 1, 2$ and $k \neq j$.

The regression of Y onto a single predictor, X_j , has slope coefficient

$$\beta_j^{(1)} = E[\ (X_j - \mu_j)^2\ ]^{-1} \cdot E[\ (X_j - \mu_j) \cdot (Y - \mu_y)\ ] = \rho_{yj}\ \sigma_y / \sigma_j\ , \qquad \{ 2.32 \}$$

for $j = 1$ or 2.

PREDICTOR SIGNS CONVENTION: Note that we can change the numerical sign of X_1 and/or of X_2 , if necessary, so that neither predictor-response marginal correlation is negative: $\rho_{y1} \geq 0$ and $\rho_{y2} \geq 0$. Thus, without loss of generality, we need only consider the case where both single-regressor coefficients are non-negative in { 2.29 }: $\beta_1^{(1)} \geq 0$ and $\beta_2^{(1)} \geq 0$.

When the response variable, Y, is regressed onto both X_1 and X_2, the resulting slope coefficients are defined by the matrix equations

$$\beta = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \rho_{y1}\sigma_y\sigma_1 \\ \rho_{y2}\sigma_y\sigma_2 \end{bmatrix},$$

which yield the coefficients

$$\beta_1 = (\rho_{y1} - \rho_{12}\rho_{y2}) \sigma_y / [\sigma_1 (1 - \rho_{12}^2)], \qquad \{ 2.33 \}$$

and

$$\beta_2 = (\rho_{y2} - \rho_{12}\rho_{y1}) \sigma_y / [\sigma_2 (1 - \rho_{12}^2)], \qquad \{ 2.34 \}$$

assuming that $|\rho_{12}|$ is strictly less than 1 (i.e. assuming the inverse matrix exists!)

WRONG SIGN PROBLEMS: We will say that a "wrong sign" problem has NOT occurred as long as both $\beta_1 \geq 0$ and $\beta_2 \geq 0$ when $\rho_{y1} \geq 0$ and $\rho_{y2} \geq 0$.

Note, from equations { 2.30 } and { 2.31 }, that "wrong sign" problems cannot occur when the predictor intercorrelation is non-positive: $\rho_{12} \leq 0$. Alas, our convention of choosing the numerical signs of X_1 and X_2 so that $\beta_1^{(1)} \geq 0$ and $\beta_2^{(1)} \geq 0$, tends (at least in my experience) to yield a numerical value for ρ_{12} that is strictly positive.

We now argue that "wrong sign" problems tend to occur when $\rho_{12} > 0$ because ρ_{y1} and ρ_{y2} are spread apart, numerically. And the numerical difference between ρ_{y1} and ρ_{y2} does NOT need to be very large at all before a "wrong sign" may be produced, at least when ρ_{12} is near its upper limit of one. For example, equations { 2.30 } and { 2.31 } show that, whenever $\rho_{12} > 0$ and $\rho_{y1} > \rho_{y2}$, then

(i) β_1 will always have a "believably" positive sign, while

(ii) β_2 may be "unbelievably" negative, and certainly will be if ρ_{12} is sufficiently close to 1.

On the other hand, the exact reverse sort of situation ($\beta_1 < 0$ and $\beta_2 > 0$) can occur when $\rho_{12} > 0$ and $\rho_{y1} < \rho_{y2}$.

To consider all possibilities, we now denote the ratio of ρ_{y1} to ρ_{y2} by $\Re = \rho_{y1} / \rho_{y2}$ and rewrite { 2.30 } and { 2.31 } as

$$\beta_1 = \sigma_y \cdot \rho_{y2} \cdot (\Re - \rho_{12}) / [\sigma_1 (1 - \rho_{12}^2)]$$

and

$$\beta_2 = \sigma_y \cdot \rho_{y2} \cdot (1 - \rho_{12} \cdot \Re) / [\sigma_2 (1 - \rho_{12}^2)] .$$

Note that the numerical sign of β_1 will agree with that of $(\Re - \rho_{12})$, while the numerical sign of β_2 will agree with that of $(1 - \rho_{12} \cdot \Re)$.

For the joint distribution of Y, X_1, and X_2 be non-singular, the determinant of the **V** matrix of equation { 2.28 } must be strictly positive. This condition can be written as $\sigma_y^2 \cdot \sigma_1^2 \cdot \sigma_2^2 \cdot (1 + 2 \cdot \rho_{y1} \cdot \rho_{y2} \cdot \rho_{12} - \rho_{y1}^2 - \rho_{y2}^2 - \rho_{12}^2) > 0$. Assuming that σ_y^2, σ_1^2 and σ_2^2 are each positive, this non-singularity condition can then be rewritten in terms of the three parameters ρ_{12}, $\Re = \rho_{y1} / \rho_{y2}$ and ρ_{y2} as...

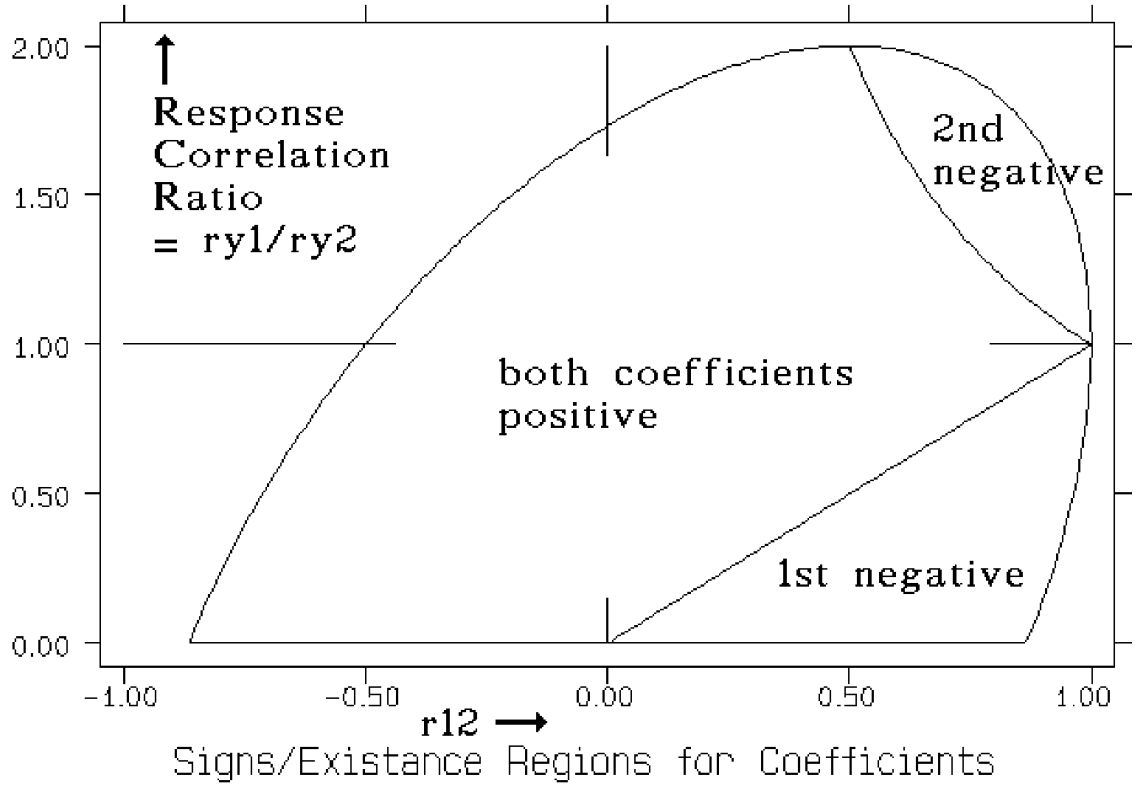
$$\max[0, \ \rho_{12} - \sqrt{(1 - \rho_{12}^2) \cdot (\frac{1}{\rho_{y2}^2} - 1)}] < \Re < \rho_{12} + \sqrt{(1 - \rho_{12}^2) \cdot (\frac{1}{\rho_{y2}^2} - 1)}$$

Figure 2.3 below illustrates how the numerical signs of β_1 and β_2 will vary, all in the special case where $\rho_{y2} = 0.5$, over subsets of the rectangular region defined by $-1 < \rho_{12} < +1$ (along the horizontal axis) and $0 \leq \Re < 1 / \rho_{y2}$ (along the vertical axis). Values outside of the upper part of the ellipse shown in Figure 2.3 are impossible in the sense that they would either violate the above tri-variate non-singularity condition or else violate the $\rho_{y1} \geq 0$ and $\rho_{y2} \geq 0$ sign convention that we adopted above.

The primary message of Figure 2.3 is that there are well defined regions in which multiple regression coefficients “should" have the “wrong" numerical sign when $\rho_{y2} = 0.5$. Of course, corresponding regions where regression coefficients either have “wrong" signs or are “undefined" also exist for all other values of ρ_{y2} over the range $0 \leq \rho_{y2} \leq 1$.

In addition to the “wrong signs" problems that can occur when $\rho_{12} > 0$, other sorts of problems are almost guaranteed to occur whenever $|\rho_{12}|$ is close to 1. In particular, note that $(1 - \rho_{12}^2)$ terms appear in the denominators of both equations { 2.30 } and { 2.31 }. Thus, numerical values for β_1 and β_2 coefficients that are quite large in absolute value are almost inevitable whenever $|\rho_{12}|$ is near one.

Figure 2.3 Signs of Coefficients in Two Predictor Regression Models when $r_{y2} = 0.5$.



Another perspective on the above observations is provided by the decomposition of β_1 and β_2 into uncorrelated γ components. For example, in the special case we are considering where $\sigma_y = \sigma_1 = \sigma_2$, principal axes again fall at 45° angles with the given predictor axes. As a result, $\beta_1 = (\gamma_1 + \gamma_2) / \sqrt{2}$ and $\beta_2 = (\gamma_1 - \gamma_2) / \sqrt{2}$, where

$$\gamma_1 = (\rho_{y1} + \rho_{y2}) / [(1 + \rho_{12}) \cdot \sqrt{2}]$$

and

$$\gamma_2 = (\rho_{y1} - \rho_{y2}) / [(1 - \rho_{12}) \cdot \sqrt{2}] .$$

Now, note that only γ_2 is sensitive to the numerical difference between ρ_{y1} and ρ_{y2} . Furthermore, only γ_2 increases in absolute size as ρ_{12} approaches +1 ; in fact, γ_1 decreases in absolute value as ρ_{12} increases! And “wrong" sign problems emerge whenever $|\gamma_2|$ exceeds γ_1 .

2.10.2 Actual Practice ...Inferences Conditional on Given Predictor Variables

Now that details of the statistical theory of two-predictor regressions have been outlined, we can explore their implications in real-life applications. In actual practice, the observed numerical value of $r_{12} = \widehat{\rho}_{12}$ (the sample estimate of ρ_{12}) may be more of an artifact of how the observed data were collected than of any sort of measure of the "natural" joint-variation of X_1 and X_2 .

> When our data come from a "designed" experiment, we have hopefully observed responses at a set of pairs of numerical values for X_1 and X_2 with the highly desirable property that $\widehat{\rho}_{12}$ is close to zero!
>
> When our data collection is merely "observational" or we have simply compiled historical (retrospective) results, the implied $\widehat{\rho}_{12}$ could be quite large. This does not imply, however, that X_1 and X_2 "should" or naturally "would" track each other ...with both tending either to increase together or to decrease together from observation-to-observation of the process under study. In other words, we have only observed responses corresponding to a strict subset of the (X_1 , X_2) pairings that would have been explored in any sort of "well-designed" experimental situation.

Conditional upon the given X_1 and X_2 predictor coordinates, $\widehat{\rho}_{12}$ is simply a known constant. But $\widehat{\rho}_{y1}$ and $\widehat{\rho}_{y2}$ remain stochastic given X_1 and X_2 . In fact, their ratio $\Re = \rho_{y1} / \rho_{y2}$ is not only stochastic given X_1 and X_2 but also numerically unstable (due to ill-conditioning) when $\widehat{\rho}_{12}$ is anywhere near to $+1$. In other words, very small numerical changes in the N observed response y-values can result in major changes in the relative-magnitudes and numerical-signs of β_1 and β_2 of equations { 2.30 } and { 2.31 }.

2.11 Tests of General Linear Hypotheses

A "general linear hypothesis" will be denoted by

$$H: \ \mathbf{A}\,\boldsymbol{\beta} = \boldsymbol{\rho} \qquad \{ 2.35 \}$$

where **A** is a known (r x P) matrix with $1 \leq \text{rank}(\mathbf{A}) = r \leq P$ and $\boldsymbol{\rho}$ is a known (r x 1) vector. When r > 1, assume also that $\mathbf{A}\,\boldsymbol{\beta} = \boldsymbol{\rho}$ are self-consistent linear equations.

Note that the equation, $\mathbf{A}\,\boldsymbol{\beta} = \boldsymbol{\rho}$, of { 2.32 } places a restriction on potential values for the $\boldsymbol{\beta}$ vector. Regression practitioners, using estimates of $\boldsymbol{\beta}$ derived from their data, can make statistical inferences about whether this restriction seems plausible. For example, when the conditional distribution of **y** given **X** is assumed to be multivariate normal, the least squares confidence region for $\mathbf{A}\,\boldsymbol{\beta}$ has the following well-known distribution and properties. Let $F(r, n-p-1; \alpha)$ denote the upper $100(1-\alpha)\%$ point of Snedecor's F-distribution, and let s^2 denote the residual mean square for error of equation { 2.22 }. Then the set of all possible vectors of the form **A b** such that

$$(\mathbf{b} - \mathbf{b}^{o})^{T} \mathbf{A}^{T} [\mathbf{A} (\mathbf{X}^{T}\mathbf{X})^{+} \mathbf{A}^{T}]^{+} \mathbf{A} (\mathbf{b} - \mathbf{b}^{o}) / (r\, s^{2}) \;\leq\; F(r, n - p - 1; \alpha) \qquad \{2.36\}$$

is an unbiased confidence region for $\mathbf{A}\boldsymbol{\beta}$ which covers the unknown true value of $\mathbf{A}\boldsymbol{\beta}$ with probability $(1 - \alpha)$ for every $\boldsymbol{\beta}$ and for every σ^2 under normal distribution theory. This region constitutes the surface and interior of an r-dimensional hyperellipsoid "centered" at $\mathbf{A}\,\mathbf{b}^{o}$, the unbiased estimate of $\mathbf{A}\boldsymbol{\beta}$.

The general solution for $\boldsymbol{\beta}$ to a set of self-consistent linear equations $A\,\beta = \rho$ can be written as

$$\boldsymbol{\beta} = \mathbf{A}^{-}\boldsymbol{\rho} + (\mathbf{I} - \mathbf{A}^{-}\mathbf{A})\,\mathbf{z} , \qquad \{2.37\}$$

where $\mathbf{A}^{-}$ is any generalized inverse for the $\mathbf{A}$ matrix and $\mathbf{z}$ is any (P x 1) vector, Rao(1973), pp. 24-25. Thus { 2.34 } will include $\boldsymbol{\beta}$ vectors outside as well as any inside the hyperellipsoid of { 2.33 }. The restricted least squares estimator of $\boldsymbol{\beta}$ under the hypothesis H : $\mathbf{A}\boldsymbol{\beta} = \boldsymbol{\rho}$ of { 2.32 } is the $\mathbf{b}$ vector (call it $\mathbf{b}^{H}$, say) that minimizes the Lagrange multiplier equation

$$\psi (\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\,\mathbf{b})^{T} (\mathbf{y} - \mathbf{X}\,\mathbf{b}) - 2\,\boldsymbol{\eta}^{T} (\mathbf{A}\boldsymbol{\beta} - \boldsymbol{\rho}) . \qquad \{ 2.38 \}$$

In other words, $\partial\,\psi / \partial\,\mathbf{b} = \mathbf{0}$ and $\partial\,\psi / \partial\,\boldsymbol{\eta} = \mathbf{0}$ together imply that

$$\begin{aligned} \mathbf{b}^{H} &= \mathbf{A}^{*}\boldsymbol{\rho} + (\mathbf{I} - \mathbf{A}^{*}\,\mathbf{A})\,\mathbf{b}^{o} , \\ &= \mathbf{b}^{o} - \mathbf{A}^{*} (\mathbf{A}\,\mathbf{b}^{o} - \boldsymbol{\rho}) , \end{aligned} \qquad \{2.39\}$$

where $\mathbf{A}^{*} = (\mathbf{X}^{T}\mathbf{X})^{+} \mathbf{A}^{T} [\mathbf{A} (\mathbf{X}^{T}\mathbf{X})^{+} \mathbf{A}^{T}]^{+}$ is one specific choice for the generalized inverse, $\mathbf{A}^{-}$, in $\mathbf{A}\,\mathbf{A}^{-}\,\mathbf{A} = \mathbf{A}$.

The resulting F-statistic for the test if the hypothesis H : $\mathbf{A}\boldsymbol{\beta} = \boldsymbol{\rho}$ of {2.32} is thus

$$\begin{aligned} F &= (\mathbf{b}^{o} - \mathbf{b}^{H})^{T}\, \mathbf{X}^{T}\mathbf{X} (\mathbf{b}^{o} - \mathbf{b}^{H}) / (r\, s^{2}) , \\ &= (\mathbf{A}\,\mathbf{b}^{o} - \boldsymbol{\rho})^{T}\, [\mathbf{A} (\mathbf{X}^{T}\mathbf{X})^{+} \mathbf{A}^{T}]^{+} (\mathbf{A}\,\mathbf{b}^{o} - \boldsymbol{\rho}) / (r\, s^{2}) \end{aligned} \qquad \{2.40\}$$

with numerator degrees-of-freedom=r and denominator degrees-of-freedom=(N – R – 1).

2.12 Weighted Residual Analyses

An extremely important phase of multiple regression modeling is that in which a practitioner examines fitted residuals to...

(i) uncover evidence of systematic lack-of-fit in the expectation model;

Examples of these sorts of problems include unrecognized “curvature" that might be removed by a well-chosen transformation of response and/or predictor variables and also failure to include certain available explanatory variables.

(ii) identify individual observations exerting undue “influence" on the fit;

Examples here include high “leverage" predictor variable combinations and/or “outlying" (maverick) response values.

and/or to

(iii) detect violations of the assumed variance-covariance structure.

Mild deviations from an assumed homoscedastic, uncorrelated error structure usually may not have any major effects. But blatant deviations from an assumed dispersion structure can make the “usual formulas" quite poor indicators of reality.

Our treatment here will avoid almost all practical aspects of residual analyses; highly readable information on basic strategy/tactics is provided by, say, Chapter 3 of Draper and Smith(1981) or Chapters 5 and 6 of Weisberg(1980) or the book of Belsley, Kuh and Welsch(1980). In fact, all we really want to do here is to display/discuss some fundamental residual analysis formulas that are slightly more general than those available elsewhere:

Here in Section §2.12, we consider a residual analysis formulation suitable for the “weighted" least-squares fits commonly used in “robust" regression and potentially useful in visual re-regression (VRR.) Of course, our weighted formulation does include ordinary (unweighted) least-squares residuals as a special case.

In Section §3.5 at the end of Chapter 3, we discusses analyses of residuals resulting from shrinkage-regression fits. There we show how shrinkage introduces bias into residuals when the expectation model is correct, just as shrinkage introduces bias into coefficient estimates. In fact, shrinkage usually changes the variance, the leverage, and the overall influence of each and every observation!

We start by considering a heteroscedastic variance formulation for multiple regression models that generalizes the homoscedastic variance case described in equation { 2.2 }. The fundamental mechanism involved in this form of “robust" fitting is to reduce the weights assigned to observations that are candidates for outliers in the sense that their fitted residuals are relatively large. We will not consider details of iterative methods for defining these weights here in Chapter 2; these topics are treated in Chapter 9. Instead, we proceed here as if observation weights are given values.

While regression weights are usually viewed as being inversely proportional to the variances of their observations, it certainly is not mandatory to visualize outliers as having high variability. Indeed, outliers can also result because their expected values deviate wildly from the general

pattern of other observations. In any case, assigning a weight of zero to an observation certainly does not imply that that observation has infinite variance. In fact, as detailed below, a weight of zero really implies simply that the expected value and the variance of that observation could be any pair of finite values. On the other hand, assigning strictly positive weights to a subset of observations will be interpreted here as an attempt to make both of the resulting model equations (expected value and variance) fit relatively well to all data points of that subset.

The (un-centered) multiple regression model considered here in Section §2.11 will be:

Conditional Expectation of $\mathbf{u}$ given $\mathbf{Z}$: $\quad E(\mathbf{u} \mid \mathbf{Z}) = \mathbf{Z}\boldsymbol{\beta}$, and $\qquad \{ 2.41 \}$

Conditional Variance of $\mathbf{u}$ given $\mathbf{Z}$: $\quad V(\mathbf{u} \mid \mathbf{Z}) = \sigma^2 \mathbf{W}^{-}$, $\qquad \{ 2.42 \}$

where $\mathbf{Z}$ is $(P+1) \times N$ [with the $\mathbf{1}$ vector as its first column] and $\mathbf{W}^{-}$ is any diagonal, generalized-inverse matrix [Rao(1973), page 24] for the $N \times N$ diagonal and non-negative definite matrix of finite weights, $\mathbf{W}$. I.E. $\mathbf{W}^{-}$ is any diagonal matrix such that $\mathbf{W}\,\mathbf{W}^{-}\mathbf{W} = \mathbf{W}$. In other words, w_{ii}^{-} must equal $1/w_{ii}$ whenever $w_{ii} > 0$, but w_{ii}^{-} can be any finite value whenever $w_{ii} = 0$ for $1 \le i \le N$.

Consider now the transformations

$$\mathbf{y}^* = \mathbf{W}^{1/2}\,\mathbf{u} \quad \text{and} \quad \mathbf{X}^* = \mathbf{W}^{1/2}\,\mathbf{Z}, \qquad \{ 2.43 \}$$

where the diagonal matrix of positive-square-roots, $\mathbf{W}^{1/2}$, is uniquely determined from $\mathbf{W}$. The rows of $\mathbf{y}^*$ and $\mathbf{X}^*$ corresponding to zero weights are thus identically zero; after all, we have explicitly excluded all cases where the corresponding rows of $\mathbf{u}$ and $\mathbf{Z}$ might contain infinite values ($\pm\infty$.)

Our generalized regression models of equations { 2.38 } and { 2.39 } thus imply the models $E(\mathbf{y}^* \mid \mathbf{X}^*) = \mathbf{X}^*\boldsymbol{\beta}$ and $V(\mathbf{y}^* \mid \mathbf{X}^*) = \sigma^2 \cdot \mathbf{D}$, where $\mathbf{D}$ is a $N \times N$ diagonal matrix each of whose diagonal elements is either a zero or a one. [Note that the $\mathbf{D}$ matrix is its own uniquely-determined Moore-Penrose inverse, Rao(1973), page 26.] Letting $N^* \le N$ denote the rank of $\mathbf{D}$, the degrees-of-freedom-for-error in our generalized model are $N^* - P - 1$, at least when $\mathbf{X}^*$ is of full rank, namely $P + 1$. The corresponding least-squares estimates of β and σ^2 from the regression of $\mathbf{y}^*$ onto $\mathbf{X}^*$ are then given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\,\mathbf{X}^{*T}\mathbf{y}^* = (\mathbf{Z}^T\mathbf{W}\,\mathbf{Z})^{-1}\,\mathbf{Z}^T\mathbf{W}\,\mathbf{u} \qquad \{ 2.44 \}$$

and

$$s^2 = (\mathbf{u} - \mathbf{Z}\widehat{\boldsymbol{\beta}})^T\mathbf{W}(\mathbf{u} - \mathbf{Z}\widehat{\boldsymbol{\beta}}) / (N^* - P - 1). \qquad \{ 2.45 \}$$

Note that $\widehat{\beta}$ is unbiased, $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and its variance is $V(\widehat{\boldsymbol{\beta}}) = \sigma^2 \cdot (\mathbf{Z}^T\mathbf{W}\,\mathbf{Z})^{-1}$, where $\sigma^2 = E(s^2)$ of { 2.42 }.

The vector of generalized residuals from model equations { 2.38 } and { 2.39 } resulting from the β and σ^2 estimates of equations { 2.41 } and { 2.42 } are, by definition, of the form

$$\widehat{\mathbf{r}}_{\mathbf{u}} \equiv \mathbf{u} - \mathbf{Z}\widehat{\boldsymbol{\beta}} = [\,\mathbf{I} - \mathbf{Q}\,\mathbf{W}\,]\,\mathbf{u} \qquad \{\ 2.46\ \}$$

where

$$\mathbf{Q} = \mathbf{Z}\,(\,\mathbf{Z}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{Z}\,)^{-1}\,\mathbf{Z}^{\mathrm{T}}\,. \qquad \{\ 2.47\ \}$$

The conditional variance-covariance matrix of $\widehat{\mathbf{r}}_{\mathbf{u}}$ given $\mathbf{Z}$ is thus of the form

$$\mathrm{V}(\widehat{\mathbf{r}}_{\mathbf{u}}\,|\,\mathbf{Z}) = \sigma^2 \cdot (\,\mathbf{W}^{-} - \mathbf{Q}\,\mathbf{D} - \mathbf{D}\,\mathbf{Q} + \mathbf{Q}\,)\,. \qquad \{\ 2.48\ \}$$

Note that two very different sorts of formulas for the ii-th diagonal element of this conditional variance matrix result when the weight given to the i-th observation is, respectively, positive or zero:

$$\mathrm{V}(\widehat{\mathrm{r}}_{\mathrm{u(i)}}\,|\,\mathrm{Z}) = \sigma^2 \cdot (\,\mathrm{w}_{\mathrm{ii}}^{-1} - \mathrm{q}_{\mathrm{ii}}\,) \quad \text{when } \mathrm{w}_{\mathrm{ii}} > 0\,, \qquad \{\ 2.49\ \}$$

but

$$\mathrm{V}(\widehat{\mathrm{r}}_{\mathrm{u(i)}}\,|\,\mathrm{Z}) = \sigma^2 \cdot (\,\mathrm{w}_{\mathrm{ii}}^{-} + \mathrm{q}_{\mathrm{ii}}\,) \quad \text{when } \mathrm{w}_{\mathrm{ii}} = 0\,. \qquad \{\ 2.50\ \}$$

Similarly, the conditional covariance between the i-th and j-th element of $\widehat{\mathrm{r}}_{\mathrm{u}}$ given Z is

$$\mathrm{Cov}(\widehat{\mathrm{r}}_{\mathrm{u(i)}}\,,\widehat{\mathrm{r}}_{\mathrm{u(j)}}\,|\,\mathrm{Z}) = -\sigma^2 \cdot \mathrm{q}_{\mathrm{ij}} \quad \text{when } \mathrm{w}_{\mathrm{ii}} > 0 \text{ and } \mathrm{w}_{\mathrm{jj}} > 0\,, \qquad \{\ 2.51\ \}$$

where $\mathrm{q}_{\mathrm{ij}} = \mathbf{z}_{\mathrm{i}}^{\mathrm{T}}\,(\,\mathbf{Z}^{\mathrm{T}}\,\mathbf{W}\,\mathbf{Z}\,)^{-1}\,\mathbf{z}_{\mathbf{j}}$ and $\mathbf{z}_{\mathrm{i}}^{\mathrm{T}}$ is the i-th row of $\mathbf{Z}$. Otherwise, this covariance is zero.

The i-th residual is standardized by dividing it by the "usual" estimate of its standard deviation, the square root of the i-th diagonal element of { 2.45 } with σ^2 estimated by s^2 of { 2.42 } :

$$\mathrm{r}^{\mathrm{s}}_{\mathrm{u(i)}} = \frac{\widehat{\mathrm{r}}_{\mathrm{u(i)}}}{\mathrm{s}\cdot\sqrt{\mathrm{w}_{\mathrm{ii}}^{-1} - \mathrm{q}_{\mathrm{ii}}}} \quad \text{when } \mathrm{w}_{\mathrm{ii}} > 0\,, \qquad \{\ 2.52\ \}$$

but

$$\mathrm{r}^{\mathrm{s}}_{\mathrm{u(i)}} = \frac{\widehat{\mathrm{r}}_{\mathrm{u(i)}}}{\mathrm{s}\cdot\sqrt{\mathrm{w}_{\mathrm{ii}}^{-} + \mathrm{q}_{\mathrm{ii}}}} \quad \text{when } \mathrm{w}_{\mathrm{ii}} = 0\,. \qquad \{\ 2.53\ \}$$

Unfortunately, the numerator and denominator of this ratio are usually not independent statistics when $\mathrm{w}_{\mathrm{ii}} > 0$; thus, standardized residuals corresponding to strictly positive weights generally do not follow Student's-t distribution under normal theory.

The i-th residual is studentized by dividing it by an independent estimate of its standard deviation. The arguments used by Beckman and Trussell(1974) and also by Ellenberg(1973,1976) are easily extended to our uncorrelated-observations, heterogeneous-variances model, equations { 2.38 } and { 2.39 }. Namely, consider the estimate of β resulting from "leaving out" the i-th observation from the model:

$$\widehat{\boldsymbol{\beta}}_{(-i)} \;=\; (\mathbf{Z}^T \mathbf{W} \mathbf{Z} - w_{ii} \cdot \mathbf{z}_i \mathbf{z}_i^T)^{-1} (\mathbf{Z}^T \mathbf{W} \mathbf{u} - w_{ii} \cdot z_i \cdot u_i) , \qquad \{ 2.54 \}$$

where $\mathbf{z}_i^T$ is the i-th row of $\mathbf{Z}$. Note that

$$\widehat{\beta} - \widehat{\beta}_{(-i)} \;=\; (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} z_i \cdot \frac{w_{ii} \cdot (u_i - \mathbf{z}_i^T \widehat{\beta})}{[1 - w_{ii} \cdot q_{ii}]} . \qquad \{ 2.55 \}$$

Thus $w_{ii} = 0$ in { 2.51 } and { 2.52 } immediately implies that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{(-i)}$. When $w_{ii} > 0$, the residual mean square, $s^2_{(-i)}$, resulting from leaving-out the i-th observation still yields an unbiased estimator of the error variance, σ^2. More importantly, this leave-out-the-i-th-observation estimate will be statistically independent of that i-th observation. Furthermore, a very simple formula for $s^2_{(-i)}$ in terms of s^2 from the full model and the i-th residual from the full model is:

$$(N^* - P - 1 - d_{ii}) \cdot s^2_{(-i)} = (N^* - P - 1) s^2 \; - \widehat{r}^2_{u(i)} / (1 - w_{ii} \cdot q_{ii}) . \qquad \{ 2.56 \}$$

In particular, $w_{ii} = 0$ implies that $d_{ii} = 0$, that the i-th residual from the regression of $\mathbf{y}^*$ onto $\mathbf{X}^*$ is zero. But $w_{ii} > 0$ implies that $d_{ii} = 1$ and that $s^2_{(-i)} \neq s^2$. The studentized residuals are thus of the general form:

$$\widehat{t}_{u(i)} \;=\; \frac{\widehat{r}_{u(i)}}{s_{(-i)} \cdot \sqrt{w_{ii}^{-1} - q_{ii}}} \;=\; r^s_{u(i)} \cdot \sqrt{\frac{N^* - P - 2}{N^* - P - 1 - [r^s_{u(i)}]^2}} \quad \text{when } w_{ii} > 0 ,$$

$$\{ 2.57 \}$$

but

$$\widehat{t}_{u(i)} \;=\; \frac{\widehat{r}_{u(i)}}{s \cdot \sqrt{w_{ii}^{-} + q_{ii}}} \;=\; r^s_{u(i)} \;\text{ when } w_{ii} = 0 . \qquad \{ 2.58 \}$$

The Cook(1977) measure of the overall influence of the i-th observation upon the regression is thus

$$\text{CINFL}_i \qquad = [\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}]^T \mathbf{Z}^T \mathbf{W} \mathbf{Z} [\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}] / (P + 1) \cdot s^2 \qquad \{ 2.59 \}$$

$$= \frac{w_{ii}^2 \cdot (u_i - \mathbf{z}_i^T \widehat{\boldsymbol{\beta}})^2 \cdot [\mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{z}_i]}{[P+1] s^2 \cdot [1 - w_{ii} \cdot \mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{z}_i]^2} .$$

Note, in particular, this influence is $\text{CINFL}_i = 0$ when $w_{ii} = 0$.

Now, defining the leverage of the i-th observation on the regression to be $\Lambda_i = 0$ when $w_{ii} = 0$ and, otherwise, to be

$$\Lambda_i = \frac{\text{Predictive Variance}}{\text{Residual Variance}} = \frac{\mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{z}_i}{[\, w_{ii}^{-1} - \mathbf{z}_i^T (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{z}_i]} \qquad \{ 2.60 \}$$

we can rewrite Cook's measure of influence as

$$\begin{aligned} \text{CINFL}_i &= 0 && \text{when } w_{ii} = 0, \\ &= \frac{(\, r^s_{u(i)})^2 \cdot \Lambda_i}{[\, P+1\,]} && \text{otherwise,} \end{aligned} \qquad \{ 2.61 \}$$

which is proportional to the product of the leverage times the square of the standardized residual.

The results derived here in Section §2.12 on analysis of weighted residuals can be summarized as follows:

When the weight given to an observation is zero, the corresponding residual can be visualized as having arbitrary variance, $V(\hat{\tau}_{u(i)} \,|\, \mathbf{Z}) = \sigma^2 \cdot (\, w_{ii}^{-} + q_{ii}\,)$ for any generalized inverse of $w_{ii} = 0$. That observation then has zero influence, zero leverage, and its studentized and standardized residuals coincide.

When the weight given to an observation is positive, the corresponding residual has variance $V(\hat{\tau}_{u(i)} \,|\, \mathbf{Z}) = \sigma^2 \cdot (\, w_{ii}^{-1} - q_{ii}\,)$ as in $\{ 2.45 \}$ and $\{ 2.46 \}$. Such an observation will have positive measures of influence and of leverage, and the corresponding studentized and standardized residuals will usually not coincide.

References for Chapter 2

Beckman, R. J. and Trussell, H. J. (1974). "The distribution of an arbitrary studentized residual and the effects of updating in multiple regression." **Journal of the American Statistical Association** 69, 199-201.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). **Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.** New York: John Wiley.

Cook, R. D. (1977). "Detection of influential observations in linear regression." **Technometrics** 19, 15-18.

Draper, N. R. and Smith, H. (1981). **Applied Regression Analysis**, Second Edition. New York: John Wiley.

Ellenberg, J. H. (1973). "The joint distribution of the standardized least squares residuals from a general linear regression." **Journal of the American Statistical Association** 68, 941-943.

Ellenberg, J. H. (1976). "Testing for a single outlier from a general linear regression." **Biometrics** 32, 637-645.

Massy, W. F. (1965). "Principal components regression in exploratory statistical research." **Journal American Statistical Association** 60, 234-256.

Obenchain, R. L. (1975a). "Residual optimality: ordinary vs. weighted vs. biased least squares." **Journal of the American Statistical Association** 70, 407-416.

Obenchain, R. L. (1975b). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics**, 17, 431-441. (Discussion: McDonald, G. C., 443-445.)

Obenchain, R. L. (1976). "Methods of ridge regression." **Proceedings of the Ninth International Biometric Conference**, Invited Papers, Volume One, 37-57, Boston.

Obenchain, R. (1977). "Classical F-tests and confidence regions for ridge regression." **Technometrics** 19, 429-439.

Rao, C. R. (1973). **Linear Statistical Inference and its Applications, 2nd edition.** New York: John Wiley & Sons.

Smith, G. and Campbell, F. (1980). "A critique of some ridge regression methods" (with discussion.) **Journal of the American Statistical Association** 75, 74-103.

Weisberg, S. (1980). **Applied Linear Regression.** New York: John Wiley.