

First...

a Preliminary Survey:

Which linear estimator is "most robust" to violations of your modeling assumptions?

$$b^{OLS} = (X'X)^{-1}X'y$$

$$b^{BLU} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

$$b^{DIAG} = (X'D^{-1}X)^{-1}X'D^{-1}y$$

$$b^{BIAS} = (X'X + kI)^{-1}X'y$$

Which linear estimator always has MSE risk optimality properties even when your assumptions are wrong?

$$b^{OLS} = (X'X)^{-1}X'y$$

$$b^{BLU} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

$$b^{DIAG} = (X'D^{-1}X)^{-1}X'D^{-1}y$$

$$b^{BIAS} = (X'X + kI)^{-1}X'y$$

Which linear estimator is always MSE risk optimal when the assumed V matrix is correct?

$$b^{OLS} = (X'X)^{-1}X'y$$

$$b^{BLU} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

$$b^{DIAG} = (X'D^{-1}X)^{-1}X'D^{-1}y$$

$$b^{BIAS} = (X'X + kI)^{-1}X'y$$

**Which linear estimator provides
MSE risk optimal RESIDUAL
estimates of lack-of-fit?**

$$b^{OLS} = (X'X)^{-1} X'y$$

$$b^{BLU} = (X'V^{-1}X)^{-1} X'V^{-1}y$$

$$b^{DIAG} = (X'D^{-1}X)^{-1} X'D^{-1}y$$

$$b^{BIAS} = (X'X + kI)^{-1} X'y$$

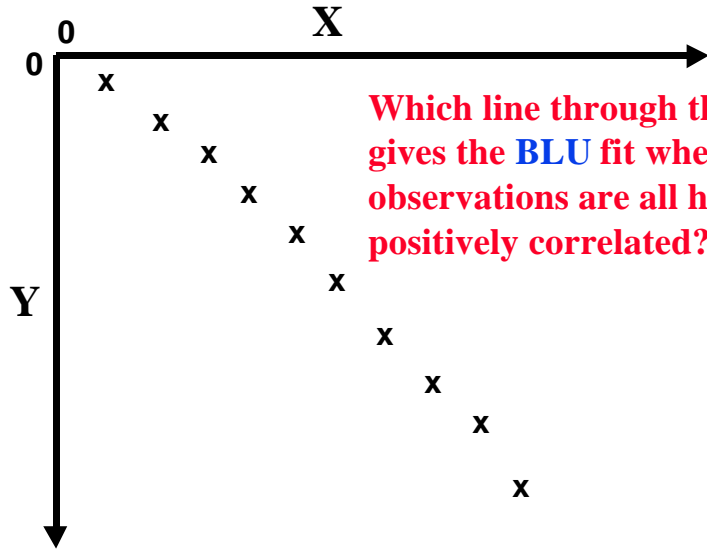
The BLU (weighted least squares) estimate has optimality properties only when the analyst's assumed X (conditional expectation) and V (variance) matrices are BOTH CORRECT. And the dominance of this estimate among unbiased, linear estimators is then STRONG. Specifically, the p-by-p variance-covariance matrix of the BLU has been minimized in the well-known Gauss-Markov sense. Namely, the difference in variance-covariance matrices (for BLU minus that for b) will be positive definite when b is linear, $b \neq BLU$, and both model matrices are correctly chosen.

Anyway, weighted least squares is clearly not always BLU because analyst's choices for X and/or V can be wrong.

Furthermore, whenever $BLU \neq OLS$, there is NO minimum mean-squared-error (MSE) risk sense in which the BLU estimate is optimal among linear estimators. Some unknown **biased** estimate always has lower MSE risk than the BLU.

But there is a sense in which OLS is ALWAYS optimal ...even when one or both parts of the assumed model are wrong!!! ...see Obenchain(1975) "Residual optimality: ordinary vs. weighted vs. biased least squares." JASA 70: 375-379. OLS produces residuals that are ALWAYS optimal estimators of lack-of-fit in the assumed conditional expectation model. The expectation, variance-covariance matrix, and MSE risk matrix of OLS residual vector are all optimal in the WEAK sense that all of their eigenvalues have been simultaneously minimized.

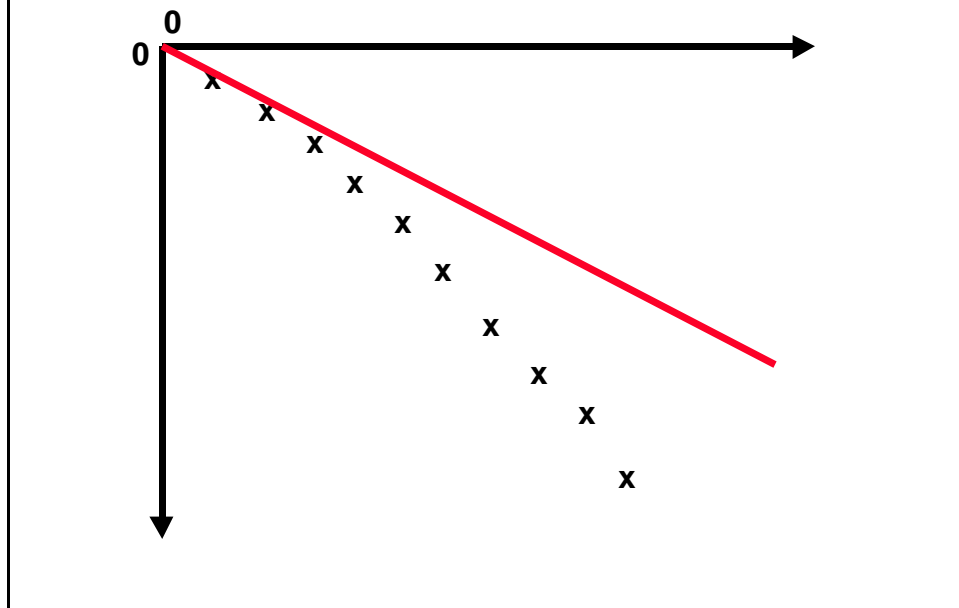
Simulated "delay" data from R. V. Laue, Bell Labs, 1971



Which line through the origin gives the BLU fit when the observations are all highly positively correlated?

I never actually saw the matrix that Dick was using for V.

Bell Labs' STATLIB / RegPak "BLU" fit...



Among the telephone transmission engineers at Bell Labs, Holmdel, Dick Laue was a real friend and supporter of statistics. Like David Goldstein, MD, of Lilly, Dick even attended some of our seminars and occasionally went to lunch with us.

Dick had attended one of the STATLIB / RegPak training sessions that AT&T and Bell Labs statisticians regularly taught at the Bell System Training Center in Lisle, Illinois. (The focus of that training was primarily on local-level forecasting of Bell System "main gain." That number would then drive all financial and engineering planning for growth.) Because of this training, STATLIB software had more than 800 regular users by 1971 ...many more than SAS at that same time. The statisticians at Bell Labs locations other than Murray Hill did not have access to S until about 1978, and there was no documentation or even a listing of S functions at that time!!! The first Bell license for SAS was purchased in about 1982.

Anyway, when Dick plotted BLU estimates from STATLIB / RegPak as shown above, he came to Bill Brelsford (primary author of the STATLIB family of Fortran programs) to ask if there could be a "bug" in the RegPak routine for BLU estimation. Anyway, Bill asked me to demonstrate that BLU estimation can indeed yield residuals all of the same sign.

NOTE: BLU estimation does produce some "optimal" residuals ...but they certainly are not the ones we can see here! [See slide 25.]

Least Squares Estimation using 2 Observations

Bob Obenchain
Eli Lilly & Company

Today's talk will explore what may well be the most simple possible "non-standard" case of least squares estimation. Suppose you have only two observations, and you assume that they have the same mean, are correlated and have different variances. Suppose you wish to compare Gauss-Markov (Best Linear Unbiased=BLU) estimation with Ordinary (unweighted) Least Squares (OLS), diagonally weighted least squares and biased estimators for this simple case.

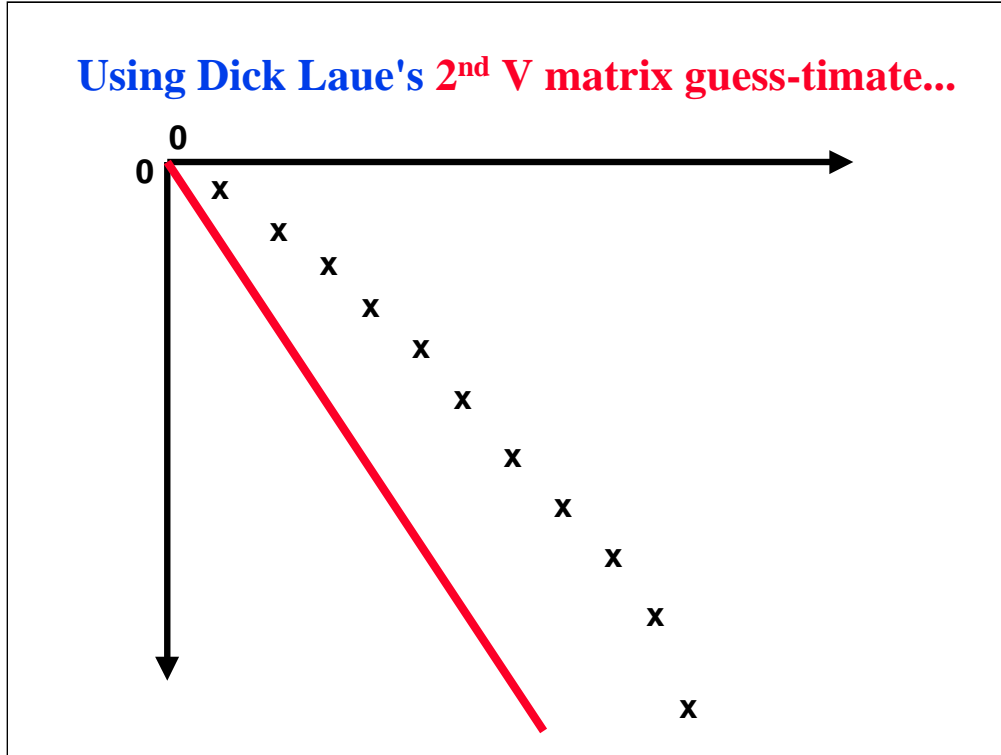
First, we argue that optimal parameter estimates do not necessarily lead to optimal predictions or residuals. For example, we show that both BLU residuals can have the same numerical sign when the assumed correlation is sufficiently positive.

Then we show that the BLU residuals are always larger than OLS residuals in a weak minimum-risk sense ...even in the limit where one BLU residual approaches zero because one observation is infinitely more precise than the other.

Obenchain(1975) showed that OLS residuals always provide optimal estimates of lack-of-fit in this weak minimum-risk sense ...even when there are more than 2 observations and both parts (expectation and dispersion) of the assumed model are wrong!

In fact, we will argue that BLU estimation NEVER provides minimum MSE risk estimates of regression coefficients while OLS estimation ALWAYS provides minimum MSE risk estimates of possible lack-of-fit of the model.

Using Dick Laue's 2nd V matrix guess-timate...



After working out most of the details for the 2-observations case (the main topic of this presentation) over the weekend, I visited Dick's lab on the next Monday and told him that "all BLU residuals definitely can have the same numerical sign ...so there probably is no error in RegPak."

And Dick said that he too had been thinking about his problem over the previous weekend, and he had convinced himself that his choice for V matrix had been quite unrealistic. In fact, he had found a "much, much better" choice for V and also had added an eleventh data point.

The BLU fit from STATLIB / RegPak now looked like THIS!!!

(I never actually saw Dick's second choice of V matrix, either.)

History...

- **John Tukey, 1974**
 - my 1st presentation to John
 - wrote letter encouraging work on generalizations
- **Geoffrey Watson & Peter Bloomfield**
 - Durbin-Watson (autocorrelation)
 - "Inefficiency" of Least Squares
- **JASA Theory and Methods 1975**
 - one journal citation in 1976, none since
 - reference in 2nd edition of Draper & Smith

Reference Materials:

Plots of Laue data and Statlib/RegPak fits on semi-log paper.

April 14, 1971: "Consultation Notes" by R. L. Obenchain

March 12, 1974: letter from J. W. Tukey.

My 1975 JASA paper was listed as a reference for Chapter 3, "Examination of Residuals," in the Second Edition (1981) of **Applied Regression Analysis** by Norman Draper and Harry Smith. Alas, there is no corresponding reference in the Third Edition (1998) ...were only entire books on residual analysis are referenced.

Assumptions, Optimality and Robustness of Estimates for the Linear Model

Multiple Linear Regression models, also known simply as "linear models," are unquestionably the most pervasively studied and commonly applied tools of statisticians and econometricians. Yet, regression practitioners have an unfortunate tendency to talk about alternative regression estimation methods using terminology that is unclear and/or potentially misleading.

For example, writers all too frequently say things like... "Ordinary least squares estimation assumes that the observations being modeled are uncorrelated and homoscedastic." (Observations are "homoscedastic" when they have a common, constant variance.) A more accurate statement would read something like... "Ordinary least squares provides Best Linear Unbiased (BLU) estimates when the observations being modeled are uncorrelated and homoscedastic."

What are the implications of this subtle distinction in actual regression practice?

Suppose that a regression practitioner has completely specified his/her model in the sense that he/she has written down a specific conditional variance-covariance matrix of responses as well as a linear model for the conditional mean response. These regression practitioners still need to decide which estimation methodology to actually apply! It is true that many practitioners simply use the estimation methodology that is commonly said to be optimal for their ASSUMED model. This is probably what their teachers expected them to do ...and this is the only option actually implemented in all known software packages.

More introspective practitioners recognize that their assumed model is probably "over-simplified" and, thus, quite possibly (if not almost surely) a WRONG model. What should they do? Which least squares estimator is "most robust" to violations of model assumptions?

Assumed Multiple Regression Model

$$E(y|X) = Xb$$

$$V(y|X) = \sigma^2 \cdot V$$

If you are a statistician or econometrician, these equations are probably much too familiar to need a detailed explanation. But, here goes!!!

The first matrix equation states that the conditional expected value of a n-by-1 vector of responses, y , given a known n-by-p matrix of predictor (covariate) values, X , lies somewhere within the column space of this X matrix (a p-dimensional subspace of n-dimensional euclidean space.) And the p-by-1 vector of unknown, true regression coefficients, β , defines the specific linear combination that represents this expectation.

The second matrix equation states that the conditional variance-covariance matrix of the n-by-1 vector of responses, y , given the n-by-p matrix of predictor (covariate) values, X , is some unknown, non-negative scalar multiple, σ^2 , of a known, positive definite (symmetric) matrix, V .

Actual Truth

$$E(y | X) = \mathbf{m}$$

$$V(y | X) = \Sigma$$

It will be quite useful in our discussion to have symbols that represent UNKNOWN, TRUE values for the conditional mean and variance of y given X .

The unknown, true conditional expected value of y given X will be denoted by the p -by-1 vector μ .

The unknown, true conditional variance-covariance matrix of y given X will be denoted by the n -by- n matrix Σ .

Knowns (observed or assumed)...

y, **X** and **V**

Unknowns...

b, **s²**, **m** and **S**

Now here's not only what we think we "know" but also what we need to admit that we "don't know"...

Estimation methods for linear models tend to focus primary attention on the regression coefficients vector, β , and secondary attention on estimation of σ .

Here, we will worry primarily about the implications of using a "wrong" V matrix.

With SAS proc MIXED, the user isn't required to "know" V . Instead, the user claims (in a REPEATED statement) that he/she knows the "TYPE=" of the proc MIXED "R" matrix, where $V = ZGZ' + R$ and $G = 0$ whenever the model contains no random effects (i.e. β is a vector of fixed effects only.) The noise TYPEs available in proc MIXED include VC or SIMPLE (uncorrelated and homoscedastic), CS = compound symmetry (exchangeable), UN = unstructured (correlated and heteroscedastic), autoregressive, spatial, Toeplitz, etc., etc.

The estimates resulting from this sort of "estimated V " are not LINEAR estimates of β (and its associated σ^2 .) Many of the results discussed here technically require V to consist of known constants.

Expectation Model Correct:

$$\mathbf{m} = X\mathbf{b} \quad \text{for some } \mathbf{b}$$

Expectation Model Incorrect:

$$\mathbf{m} \neq X\mathbf{b} \quad \text{for any } \mathbf{b}$$

Dispersion Model Correct:

$$\Sigma = \mathbf{s}^2 \cdot V \quad \text{for some } \mathbf{s}^2$$

Dispersion Model Incorrect:

$$\Sigma \neq \mathbf{s}^2 \cdot V \quad \text{for any } \mathbf{s}^2$$

**All Models are Wrong;
Some are Useful.**

G.E.P. Box (1975)

Statisticians should never forget this key point.

Econometricians think that they "know" their assumed (highly local) models are CORRECT because the form of the model was dictated by an ECONOMIC THEORY (and the order of approximation chosen.) After all, unbiasedness is a desirable property in estimation ONLY when one's model is correct. And that's essentially their definition of the "scientific method."

Statisticians should feel the need to be a little bit more introspective than econometricians.

Only OLS estimates are ALWAYS optimal ...even when the assumed linear model is wrong.

In other words, OLS is the form of least squares estimation that is most robust to violations of model assumptions.

Linear Estimators of Regression Coefficients...

$$\mathbf{b} = \mathbf{L}\mathbf{y}$$

where \mathbf{L} is a ($p \times n$) matrix of known constants that may depend upon \mathbf{X} and \mathbf{V} but not upon \mathbf{y} (nor upon \mathbf{b} , \mathbf{m} , Σ .)

Here's the well-known definition for a "linear estimator" of β .

Ordinary Least Squares:

$$b^{\circ} = X^{+} y$$

Moore-Penrose
Inverse

$$= (X' X)^{-1} X' y$$

when X is of "full" (column) rank.

The "ordinary" (or unweighted) least squares (OLS) estimator of β is the linear estimator with "o" as its superscript ...and is defined by the above extremely "familiar" equations.

Note that, instead of "assuming" that $V = I$, OLS simply "ignores" the assumed form of V .

Gauss-Markov (BLU) Estimator:

$$\hat{b} = \left(X'V^{-1}X \right)^{-1} X'V^{-1}y$$

Optimally Weighted Least Squares

The "optimally" weighted least squares estimator of β is the linear estimator with a "hat" above its symbol.

This "hat" may well cause too many statisticians and econometricians to snap to attention and salute! This estimator is Best Linear Unbiased

- NOT ONLY when the true μ lies strictly within the column-space of X
- BUT ALSO ONLY when the true Σ actually is a multiple of the assumed V matrix.

Diagonally Weighted Least Squares:

$$b^D = \left(X' D^{-1} X \right)^{-1} X' D^{-1} y$$

$$\text{where } D = \text{diag}(V)$$

Here's another option available to regression practitioners.

Here we use only the diagonal elements of the assumed V to estimate β . And we denote this linear, unbiased estimator using a "D" superscript.

Biased Estimators...

$$b^* = (X'X + k \cdot I)^{-1} X' y$$

“ridge” estimator of
Hoerl-Kennard(1970)

And biased estimators (with a "*" superscript) are also available.

When the columns of X are highly correlated, the conditional expectation model is said to be "ill-conditioned." Some such cases are highly favorable to "shrinkage" in the sense that a biased estimator with greatly reduced variance can have much lower risk (expected mean-squared-error loss) than any unbiased estimator.

But even in "well-conditioned" cases, the MSE risk of very-slightly-shrunken estimators (e.g. James-Stein-like estimators) is strictly smaller than that of BLU estimates of β ...even when the specified model is absolutely correct. In this sense, BLU estimates NEVER achieve minimum MSE risk ...because they are UNBIASED (rather than shrunken.)

Optimal what?

- **Parameter estimates**
- **Predictions**
- **Residuals**

In exploratory analyses, properties of fitted residuals are of primary importance.

All sorts of numerical changes in coefficient estimates may still produce essentially the same predictions and residuals for the AVAILABLE observations. This is almost the "definition" of ill-conditioned (nearly multicollinear) regression models.

**Aren't these quantities
very closely LINKED?**

Coefficients: $\mathbf{b} = \mathbf{L}\mathbf{y}$

Predictions: $\mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{L}\mathbf{y}$

Residuals: $\mathbf{y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{X}\mathbf{L})\mathbf{y}$

How can one \mathbf{b} be optimal for estimating β and yet not optimal for estimating residuals?

The above equations aren't wrong. But things are not this simple ...either in regression theory or in actual practice.

The most intuitive explanation I can think of is the following. Biased "shrinkage" estimates can reduce the risk in estimating β via a variance-bias trade-off. Since the OLS residual vector is already as "short" as possible, no reduced (overall) variability can result from introducing biases into it.

Suppose that $T'T = V^{-1}$

Then

$$E(Ty | X) = TXb$$

And

$$V(Ty | X) = s^2 \cdot I$$

Consider the following well-know "transformational" motivation for Gauss-Markov (BLU) estimation...

Let T be any n-by-n "square root" matrix for V^{-1} in the sense that $T'T = V^{-1}$.

T is not uniquely determined by this requirement; if H is any orthogonal matrix, then HT is another possible choice for T.

However, each of these choices for a T square-root matrix is clearly of the form

$T' = V^{-1}T^{-1}$. As a direct result, TVT' is always an n-by-n identity matrix!!!

In other words, Gauss-Markov (BLU) estimation of β on the original data is equivalent to OLS estimation of β on the transformed data, Ty.

On the other hand, residuals and predictions for the original model are NOT SIMPLY related to the residuals and predictions for the transformed model because T usually is not a diagonal matrix (for any choice of H, above.)

In other words, all such T transformations DESTROY THE IDENTITY OF THE ORIGINAL OBSERVATIONS AND RESIDUALS.

In fact, the β regression coefficient vector and the σ^2 error variance are the ONLY things these two sets of models have in common!

Two Observations...

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \& \quad \mathbf{V} = \begin{bmatrix} 1 & \mathbf{r} / \mathbf{w} \\ \mathbf{r} / \mathbf{w} & 1/\mathbf{w}^2 \end{bmatrix}$$

Model assumes 2 things:

$$\mathbf{b} = \mathbf{m}_1 = \mathbf{m}_2$$

observations are correlated
and heteroscedastic

Note that the condition for \mathbf{V} to be positive definite here is simply that $(1-\rho^2)/w^2 \neq 0$. This means that $0 < w < +\infty$ and $-1 < \rho < +1$.

Note that the relative variance of the first observation is 1 while that of the second observation is $1/w^2$.

Equivalently, the relative weight (precision) of the first observation is 1 while that of the second observation is w^2 .

In other words, as the relative weight (precision) of the second observation increases, its variance relative to the first observation decreases!

And ρ represents the correlation between the two observations.

Note that our assumed form for the above \mathbf{V} matrix involves no loss of generality ...all possible variance-covariance matrices for 2 observations are of this general $\Sigma = \sigma^2\mathbf{V}$ form for some values of σ^2 , ρ and w .

In the next 7 slides, we will NOT even worry about using incorrect values for ρ and w . Instead, we will show that OLS produces "better" residuals than BLU estimation even when the assumed ρ and w are correct!

$$b^o = \frac{y_1 + y_2}{2}$$

$$\hat{b} = \frac{(1 - rw)y_1 + w(w - r)y_2}{1 - 2rw + w^2}$$

$$b^D = \frac{y_1 + w^2 y_2}{1 + w^2}$$

It's really easy to derive these three linear, unbiased estimators of β (OLS, BLU and Diagonally Weighted) as soon as one realizes that...

$$V^{-1} = \frac{1}{(1 - r^2)} \begin{bmatrix} 1 & -rw \\ -rw & w^2 \end{bmatrix}$$

$$\hat{b} = b^D = b^o$$

when $w = 1$

Thus, in the following, we assume not only that $w \neq 1$ but also that $y_1 \neq y_2$ so that fitted residuals will not both be zero.

Note that the three unbiased estimators of β are equivalent when $w = 1$.

Similarly, the 2 observations also need to be different (numerically) so that both residuals cannot be zero.

Actually, we will focus on cases where $w \neq 1$ and $\rho \neq 0$...so that all 3 alternative estimators will be different.

BLU residuals...

$$\hat{r}_1 = w(r - w)(y_1 - y_2) / (1 - 2rw + w^2)$$

$$\hat{r}_2 = (1 - rw)(y_1 - y_2) / (1 - 2rw + w^2)$$

Both residuals will have the same numerical sign when $(r - w)(1 - rw) > 0$.

**i.e. $r > w$ when $w < 1$ or
 $r > 1/w$ when $w > 1$**

Here it is clear that both residuals will be zero when $y_1 = y_2$. Thus we retain focus on cases where $y_1 \neq y_2$.

Gauss-Markov residuals will have the same numerical sign when their numerator "factors" involving ρ have the same sign. This simply means that the product of these two factors is strictly positive.

In particular, note that Gauss-Markov residuals will never have the same numerical sign when ρ is NEGATIVE.

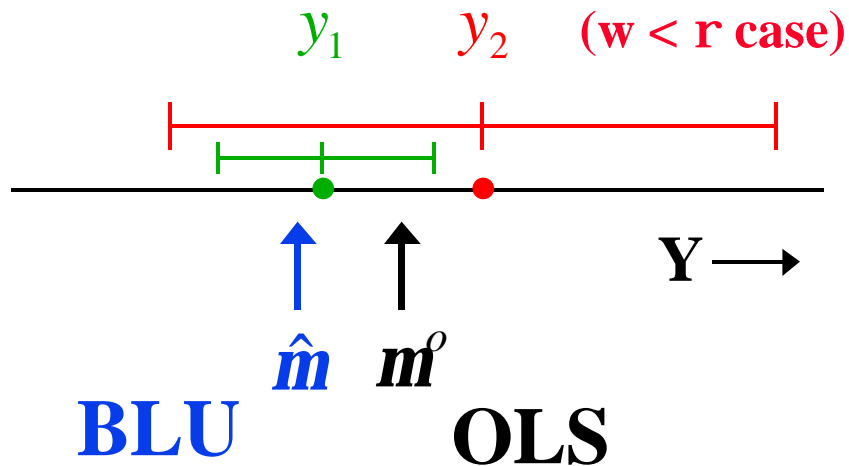
Both Gauss-Markov residuals can have the same numerical sign...

- because the observations are **heteroscedastic** and highly **positively correlated**.
- the observation with the lower variance (higher weight) will be closer to the estimated mean.

Intuitively, all this actually makes very good sense.

Still, regression practitioners need to ask themselves... "Why would I (or anybody) ever make such extreme assumptions now that I (or they) know what their ultimate implications are about fitted residuals?"

Two Positively Correlated Observations with Unequal Variances (Heteroscedastic)



Here's how to visualize the $w < \rho$ case.

The green and red horizontal "ranges" shown above simply depict the relative uncertainty (standard deviations) of the two observations.\

The ordinary (unweighted) least squares estimate will always coincide with the mid-point between the two observations.

Here the relatively uncertain observation is the one to the RIGHT and, thus, the Gauss-Markov (BLU) point estimate is to the LEFT of the more certain observation!

In other words, both residual "error terms" are much more likely to be POSITIVE here.

Both true errors would be negative here if the true mean was to the right of the less precise observation. But the uncertainty bars do NOT overlap to the right of the second observation. The uncertainty bars DO overlap to the left of the first observation ...and that is why the true mean is more likely to be there!

Residual Variance Matrices:

$$V(r^o) = \frac{\mathbf{s}^2(1 - 2rw + w^2)}{4w^2} \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}$$

$$V(\hat{r}) = \frac{\mathbf{s}^2}{w^2(1 - 2rw + w^2)} \begin{bmatrix} w^2(\mathbf{r} - w)^2 & w(\mathbf{r} - w)(1 - rw) \\ w(\mathbf{r} - w)(1 - rw) & (1 - rw)^2 \end{bmatrix}$$

Corresponding eigenvalues:

$$0 \text{ and } \mathbf{s}^2(1 - 2rw + w^2) / 2w^2 > 0$$

$$0 \text{ and } \mathbf{s}^2[(1 - rw)^2 + w^2(\mathbf{r} - w)^2] / w^2(1 - 2rw + w^2) > 0$$

Here is a exercise for the "serious student": Convince yourself that the above formulas are correct by deriving them!

$$\begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix} \begin{pmatrix} -b \\ a \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = (a^2 + b^2) \begin{pmatrix} a \\ b \end{pmatrix}$$

All two-by-two matrices of the above form with $a \neq 0$ or $b \neq 0$ are singular of rank one ...i.e. non-negative definite rather than positive definite.

This is quite easily demonstrated by (1) "guessing" the functional form for the eigenvectors of these two-by-two matrices and (2) using matrix multiplication to find the corresponding eigenvalues.

For example, we see that the column vector with $-b$ in its first row and $+a$ in its second row is (proportional to) an eigenvector with an eigenvalue of zero.

Similarly, the column vector with $+a$ in its first row and $+b$ in its second row is (proportional to) an eigenvector with an eigenvalue of (a^2+b^2) ...which is also the trace of the matrix (sum of eigenvalues) because the smaller eigenvalue is zero.

Key Observation:

$$\hat{I}_{\max} - I_{\max}^o = s^2(1-w^2)^2 / [2w^2(1-2rw+w^2)]$$

Note that this difference is **strictly greater than zero** whenever $w \neq 1$
(but $0 < w < +\infty$ and $-1 \leq \rho \leq +1$.)

Whenever OLS residuals are different from BLU residuals, OLS residuals always have lower "Total Variance."

This observation is STRIKING!

The trace of the BLU residual variance matrix always exceeds the trace of the OLS residual variance matrix even when all of the assumptions behind BLU estimation are correct i.e. even when $w \neq 1$ and $\rho \neq 0$.

On the other hand, the difference in variance matrices $V(\hat{r}) - V(r^o)$ is **not** positive definite (strong dominance)...

- In the limit as w increases to $+\infty$, the second **GM** residual approaches 0 and has variance 0, whereas the corresponding **OLS** residual has variance $\sigma^2/4$ in this limit.
- In the limit as w decreases to 0, the first **GM** residual approaches 0 and has variance $\sigma^2\rho^2$, whereas the corresponding **OLS** residual has infinite limiting variance.

For a symmetric matrix to be positive definite, it is essential for its diagonal elements to be positive. And the difference in residual variance matrices (Gauss-Markov minus OLS) does NOT have this property.

In any limit where one of the 2 observations becomes much more precise than the other, the optimally weighted least squares (BLU) estimate of their common mean must approach the more precise observation. The corresponding residual must approach zero and will have minimal asymptotic variance!

Thus the OLS residual variance matrix is **not** smaller than the BLU residual variance matrix in the familiar "**strong**" sense.

However, it is smaller in the "**weaker**" but **always applicable** sense that the strictly positive eigenvalue of the OLS residual matrix is smaller than the corresponding eigenvalue of the BLU residual variance matrix.

Thus OLS residuals dominate Gauss-Markov residuals in a "weak" sense ...but one that is always applicable.

All well-known "overall" measures of the size of a variance or MSE risk matrix are monotonically increasing functions of their eigenvalues. For example, trace = sum of eigenvalues, determinant = product of eigenvalues, squared norm = sum of squares of eigenvalues, etc., etc.

In other words, the variance-covariance (or risk) matrix of the OLS residual vector is always smaller than or equal to the variance-covariance (or risk) matrix of Gauss-Markov residuals in all of these "overall" senses.

To more fully appreciate why these sorts of results hold, we will need some more notation.

Here $V(Ty | X) = s^2 \cdot I$

for $Ty = \begin{bmatrix} (cy_1 + sy_2) / \sqrt{l_1} \\ (-sy_1 + cy_2) / \sqrt{l_2} \end{bmatrix}$

where $c^2 + s^2 = 1$ and

$cs \neq 0, l_1 \neq l_2$ when $r \neq 0$

Here are the 2 observations that Gauss-Markov (BLU) estimation is fitting and providing optimal residuals for!

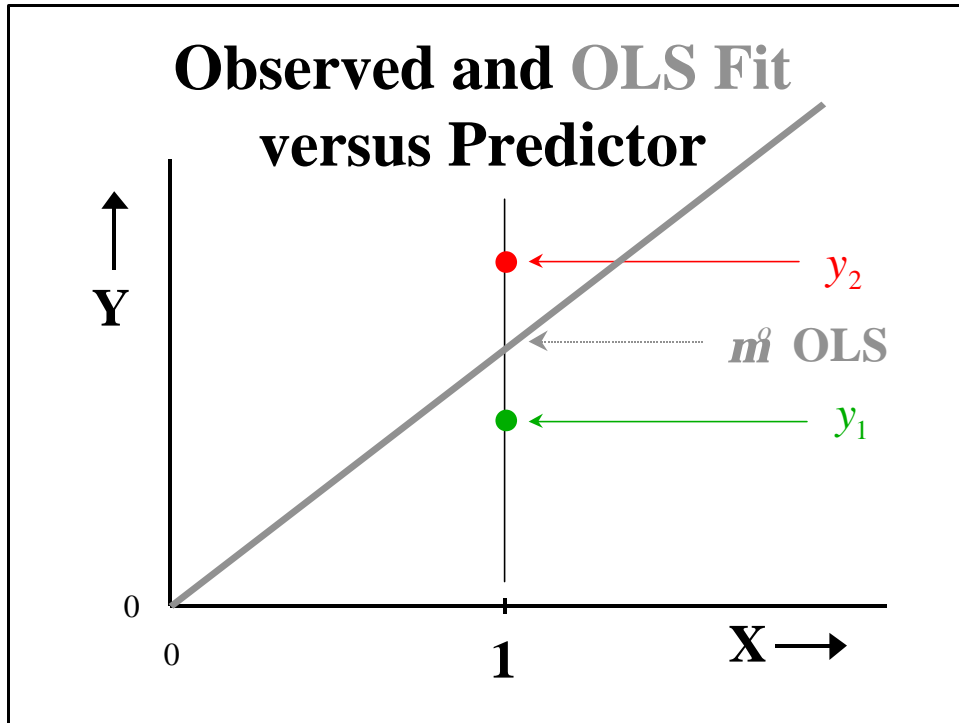
Again, this T is not uniquely determined; if H is any 2x2 orthogonal matrix of additional sines and cosines, like

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

then HT is another possible choice for T.

Again, residuals and predictions for the original model are NOT SIMPLY related to the residuals and predictions for the transformed model because HT is not a diagonal matrix for any choice of H.

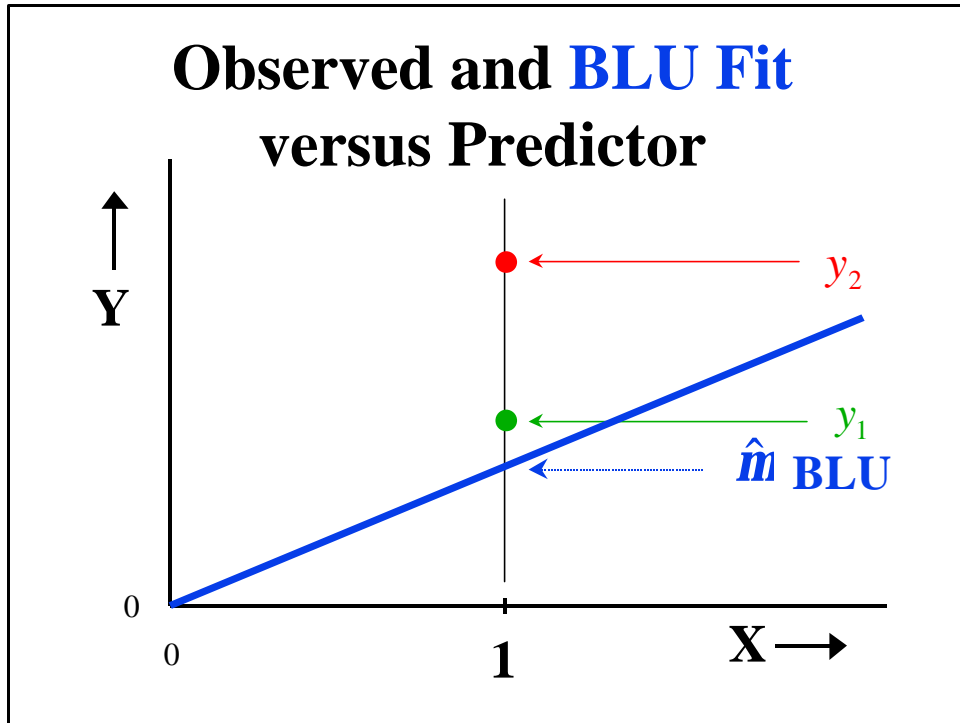
In other words, all such T transformations DESTROY THE IDENTITY OF THE TWO ORIGINAL OBSERVATIONS AND RESIDUALS.



Here we see our "2 observations" case in the traditional (Y versus X) plot for displaying the regression of a response variable onto a single predictor variable. And we have displayed the OLS fit in this graphic.

This is a really DULL graphic ...there are only 2 observations here and both occur at the SAME X value!

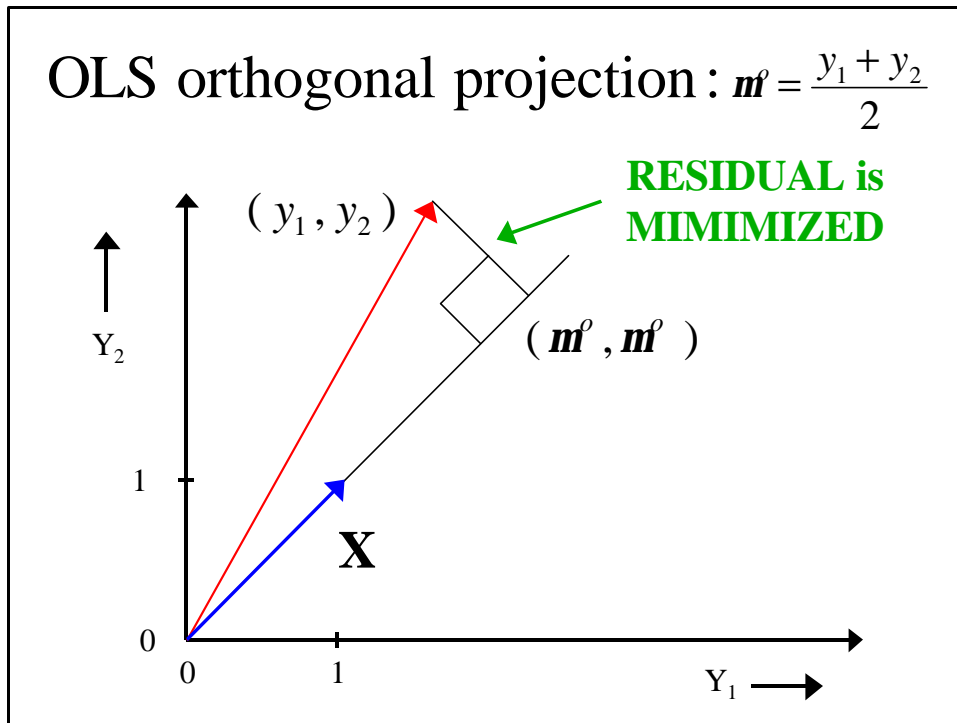
This sort of graphic is of common interest to regression practitioners simply because it is always 2-dimensional (the Y-coordinate versus the X-coordinate for each of "n" observations.)



Here, we have displayed the BLU fit for the case where the smaller (first) observation has much higher precision (smaller variance) than the larger (second) observation.

The plot that regression practitioners would really like to be able to literally "see" is the projection of the Y-vector onto the vector space spanned by the columns of X. But not only is the dimension (n) of this vector space frequently much too large to "visualize", the column space of the X matrix is also usually larger than 3 dimensions.

However, in our 2-observations case, WE CAN SEE THIS PLOT because Y is only 2-dimensional and the column space of X is 1-dimensional!!!



Here we can literally SEE the most fundamental characterization of OLS estimation ...because the relevant vector spaces are of dimension 1 and 2.

OLS is defined by the **ORTHOGONAL PROJECTION** of the observed response vector onto the known column space of X . It is easily shown [see Rao(1973), pages 46-47] that orthogonal projection is a linear operation involving a uniquely determined matrix that is both symmetric and idempotent. Here I simply use the notation we all learned in high school geometry to indicate that the vector of predictions and the vector of residuals meet at a 90 degree angle.

In other words, the very definition of (ordinary or "unweighted") **LEAST SQUARES** is that the length of the residual vector is thereby minimized. While it is obvious that the length of the predictions vector corresponding to some "oblique projection" could be either longer or shorter than the OLS prediction vector, it is also intuitively clear that "orthogonal projection" minimizes the length of the residuals vector.

References:

Rao CR. **Linear Statistical Inference and Its Applications**, Second Edition, New York: John Wiley and Sons, Inc., 1973.

Definition: $Q \equiv I - XL$

The residual vector
corresponding to $\mathbf{b} = \mathbf{L}\mathbf{y}$
can then be written as...

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{Q}\mathbf{y}$$

The Q matrix for OLS estimation is the uniquely determined orthogonal projection matrix (i.e. symmetric and idempotent) for the n-p dimensional linear subspace of n-dimensional euclidean space that is orthogonal to the columns space of X.

The *lack-of-fit* in the
expectation model

$$\mathbf{m} = \mathbf{X}\mathbf{b}$$

that can be estimated using
 $\mathbf{b} = \mathbf{L}\mathbf{y}$ is then...

$$E(\mathbf{r} | \mathbf{X}) = \mathbf{Q}\mathbf{m}$$

In the 2-observation example we considered earlier, we assumed that there was no lack-of-fit.

But worries about possible lack-of-fit in the assumed model are the primary reasons why regression practitioners "examine" fitted residuals.

Residual Optimality of OLS:

The choice $Q = Q^\circ$ simultaneously minimizes all of the eigenvalues of QAQ' for all symmetric matrices A when Q is restricted to be of the form $Q = I - XL$ (includes cases where estimates are biased and models are wrong)

Proof: Poincare Separation Theorem
Obenchain (1975) *JASA*

Note that $Q^\circ = I - \mathbf{X}\mathbf{X}^+$ has a (non-unique) decomposition $Q^\circ = \mathbf{B}\mathbf{B}'$ in which the $n \times (n-p)$ matrix \mathbf{B} is semi-orthogonal, $\mathbf{B}'\mathbf{B} = \mathbf{I}$.

Now, since $Q^\circ Q = Q^\circ$ when $Q = I - \mathbf{X}\mathbf{L}$, it follows that $Q^\circ Q A Q' Q^\circ = Q^\circ A Q^\circ$ and that the $n-p$ largest eigenvalues $Q^\circ A Q^\circ$ and of $\mathbf{B}' Q A Q' \mathbf{B}$ are equal.

The Poincare Separation Theorem [see Rao(1973), page 64] then shows that each eigenvalue of $Q A Q'$ cannot be smaller than the corresponding eigenvalue of $Q^\circ A Q^\circ$.

The choices for A of primary interest to us correspond to

- (i) $E(\mathbf{r}|\mathbf{X})'E(\mathbf{r}|\mathbf{X}) = \mathbf{m}'\mathbf{Q}'\mathbf{Q}\mathbf{m} = \text{trace}(\mathbf{Q}\mathbf{m}\mathbf{m}'\mathbf{Q}')$,
- (ii) $D(\mathbf{r}|\mathbf{X}) = \mathbf{Q}\mathbf{S}\mathbf{Q}'$, and
- (iii) $E(\mathbf{r}\mathbf{r}'|\mathbf{X}) = \mathbf{Q}(\mathbf{S} + \mathbf{m}\mathbf{m}')\mathbf{Q}'$,

for every true $\mathbf{m} = E(\mathbf{y}|\mathbf{X})$ and for every true $\mathbf{S} = D(\mathbf{y}|\mathbf{X})$.

References:

Rao CR. **Linear Statistical Inference and Its Applications**, Second Edition, New York: John Wiley and Sons, Inc., 1973.

Obenchain RL. "Residual Optimality: Ordinary Vs. Weighted Vs. Biased Least Squares." **Journal of the American Statistical Association** 70: 375-379, 1975.

“Primary” Objective?

**Optimal (BLU) estimation
of model parameters.**

**Assure that the fitted model
is appropriate for the data.**

Regression (linear model) practitioners strike me as being much too fixated on BLU estimation of regression coefficients for a model that may possibly be totally unrealistic!

The model most appropriate for your data may not provide a clear-cut answer to your research question ...but it can be a much better summary of actual reality!

Let us now examine a concrete illustration of exactly this !!!

A Convenient Formulation:

$$y_i = \mathbf{m} + z_i^* \mathbf{g} + x_i' \mathbf{b} + \mathbf{e}_i$$

where $z_i^* = 0$ or 1 treatment indicator

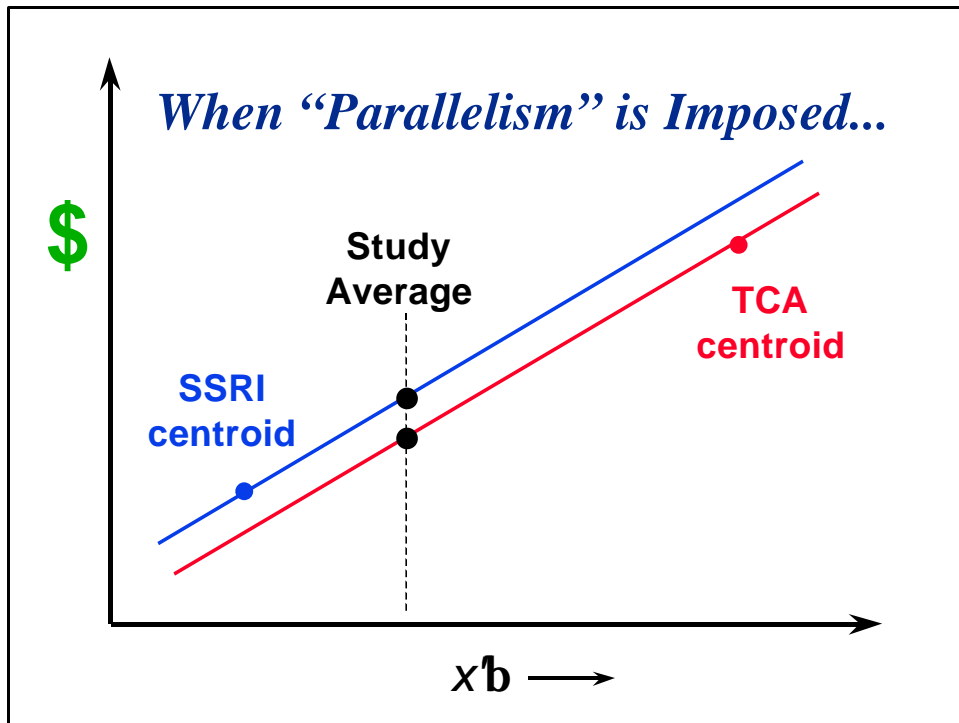
$x_i =$ covariates for i -th patient

This commonly used model for responses to two treatments adjusted for covariates assumes a "parallelism" that can be quite unrealistic.

Specifically, it assumes that all covariates have the same regression coefficients within treatment cohorts. In other words, the model assumes that treatment effects ONLY intercept terms.

It is really easy to estimate and test for this sort of effect.

Is that why this potentially quite naïve model is so widely used?



The above plot represents a situation I actually faced in analyzing some "observational" (retrospective) health care claims data.

Note that the average cost of patients on SSRIs was much lower than that of patients on TCA anti-depressants. But the (naïve) model predicted uniformly increased costs.

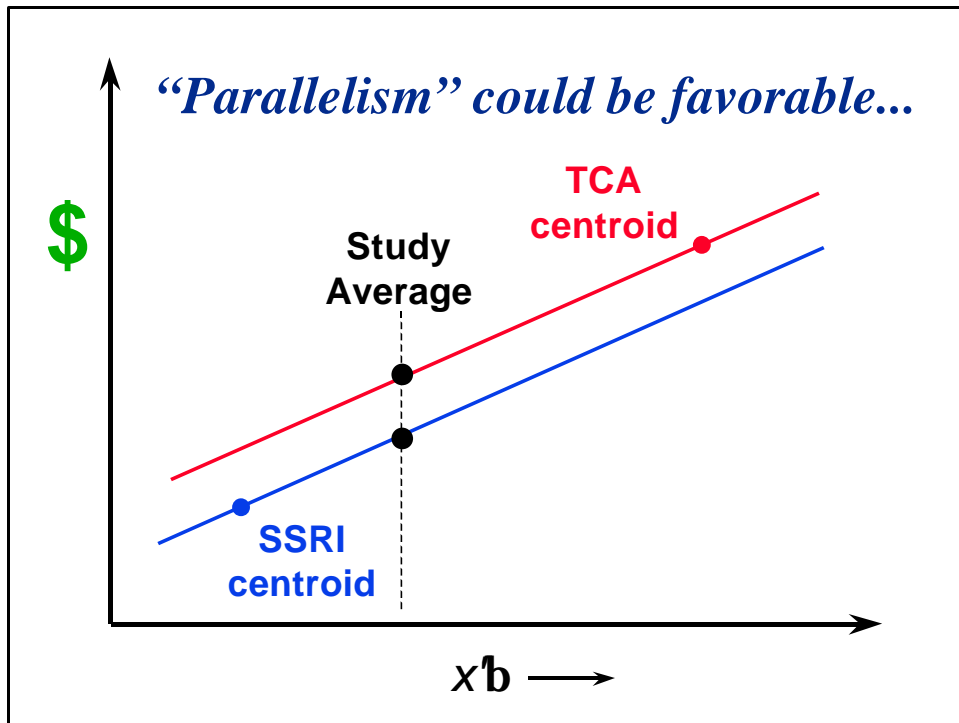
Clinical statisticians might easily say something like...

This isn't going to happen to ME!!!

Randomization will assure that both treatment groups will have approximately the same covariate centroid.

My model averages will therefore mimick the observed averages.

As we shall see, they are WRONG about this.



If this "possibility" had emerged instead, would anybody have "looked this gift-horse in the mouth?"

Quite possibly, NO!!!

But this simple "twist of fate" does not make the naïve model any more realistic.

Small data "anomalies" (like outliers or the observed positions of relatively precise and imprecise observations) can result in this model outcome instead of the "unfavorable" one on the previous slide!!!

A Realistic Formulation:

$$y_{1i} = \mathbf{m}_1 + x'_{1i} \mathbf{b}_1 + \mathbf{e}_i$$

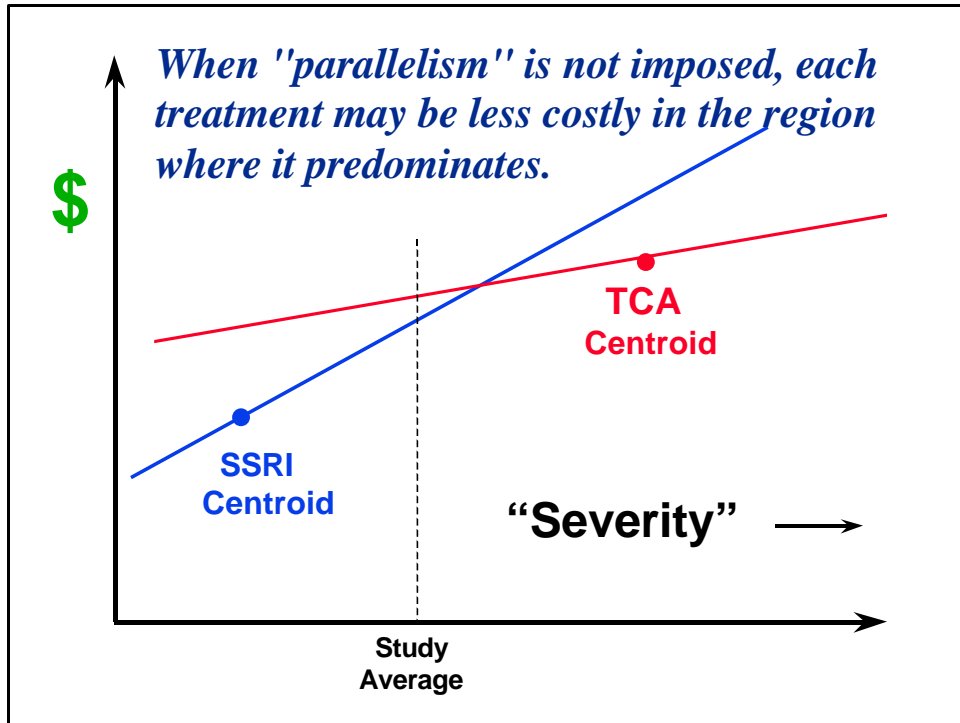
$$y_{2i} = \mathbf{m}_2 + x'_{2i} \mathbf{b}_2 + \mathbf{e}_i$$

where x_{ji} = covariate values for the i -th subject on treatment j .

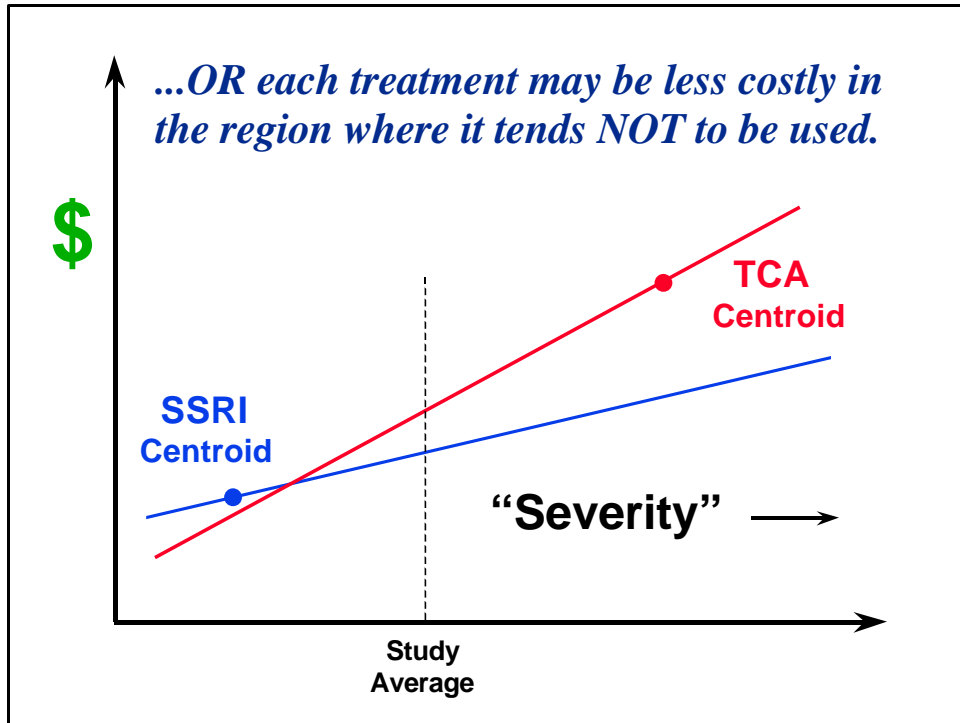
In reality, instead of being parallel, the corresponding response hyperplanes may actually intersect somewhere.

Which treatment is "better" and by "how much" is then highly dependent upon where (at which combination of covariate characteristics) response predictions are to be made!

Here we write a generalized model as if it were an econometric "two equation" system. Statisticians may well prefer to think of this phenomenon as requiring "interaction" terms (between treatments and covariates) to be added to a linear model.



But the "observational" study I analyzed did NOT predict outcomes of this form.



The "observational" study I analyzed predicted outcomes of this form.

And clinical statisticians may now be saying...

Could this possibly happen to ME ?

And my answer is... **YES !!!**

Prediction One:

The upcoming age of "genomics" will force the following REALITY upon clinical trial methodology...

Different patients respond differently to different treatments. (Results from unrealistic models become irrelevant.)

Baseline measures of human genomics will force clinical statisticians into "p over n" modeling situations. There will be great pressure upon Kerry to publish his "Bemisly Biased Information Criterion" (BBIC) work so that it can be used in clinical trials. Treatment-gene interactions will be RAMPANT!!!

To achieve any semblance of "balance," genomic information would need to be used to match or group patients before randomization. Attempts to account for gene differences after randomization will yield exactly the sorts of situations depicted in the last two slides. After all, randomization only "works" asymptotically. And 500 patients distributed "randomly" throughout a space consisting of thousands and thousands of dimensions is an incredibly small (non-asymptotic) sample!!!

[Actually, the first thing that will happen is that a relatively small number of genomic "summary measures" will emerge. And treatment interactions will initially be documented along these (relatively few) dimensions.]

Parallelism models will seem absurd! And not simply because the corresponding AIC or BIC or BBIC is relatively unfavorable. NO!!! RESIDUAL PLOTS will show that they are completely UNREALISTIC. [After all, residuals themselves are much more versatile measures of lack-of-fit than any "overall" measure of residual-mean-square error.]

Prediction Two:

Software developers will be forced into providing...

- an ever widening selection of alternative (non-"optimal") estimation methods
- more realistic estimates of the uncertainty in these alternative estimates. (i.e. OLS does NOT ASSUME observations are uncorrelated and homoscedastic.)

When you really do think that your data correspond to a certain V matrix [i.e your observations are neither homoscedastic nor uncorrelated), your statistical software should perform the analysis most appropriate for this situation.

This is especially true when you (insightfully) decide to use OLS instead of BLU estimation in this context!!!

"Introspective" statisticians realize that OLS estimation DOES NOT ASSUME that $V = I$. Historically, software developers have consistently made this curious assumption / assertion.

Prediction Three:

Bob Obenchain is much more likely to be remembered many years from now for his **Health Outcomes work at Lilly** than for his incredibly insightful and uniquely "different" contributions to the theory of **optimal / robust estimation in linear models (JASA 1975.)**

Given my current track record, I am clearly most likely to have been **COMPLETELY FORGOTTEN** "many years from now."

**All Models are Wrong;
Some are Useful.**

G.E.P. Box (1975)

As we now finish up, we find ourselves back at slide #17 (page 14.)

Statisticians really do need to be rather introspective, and empirical, and willing to "learn" from the data they analyze.

**Introspective statisticians know that only OLS estimates
ALWAYS have optimal lack-of-fit ...even when their
assumed linear model is wrong.**