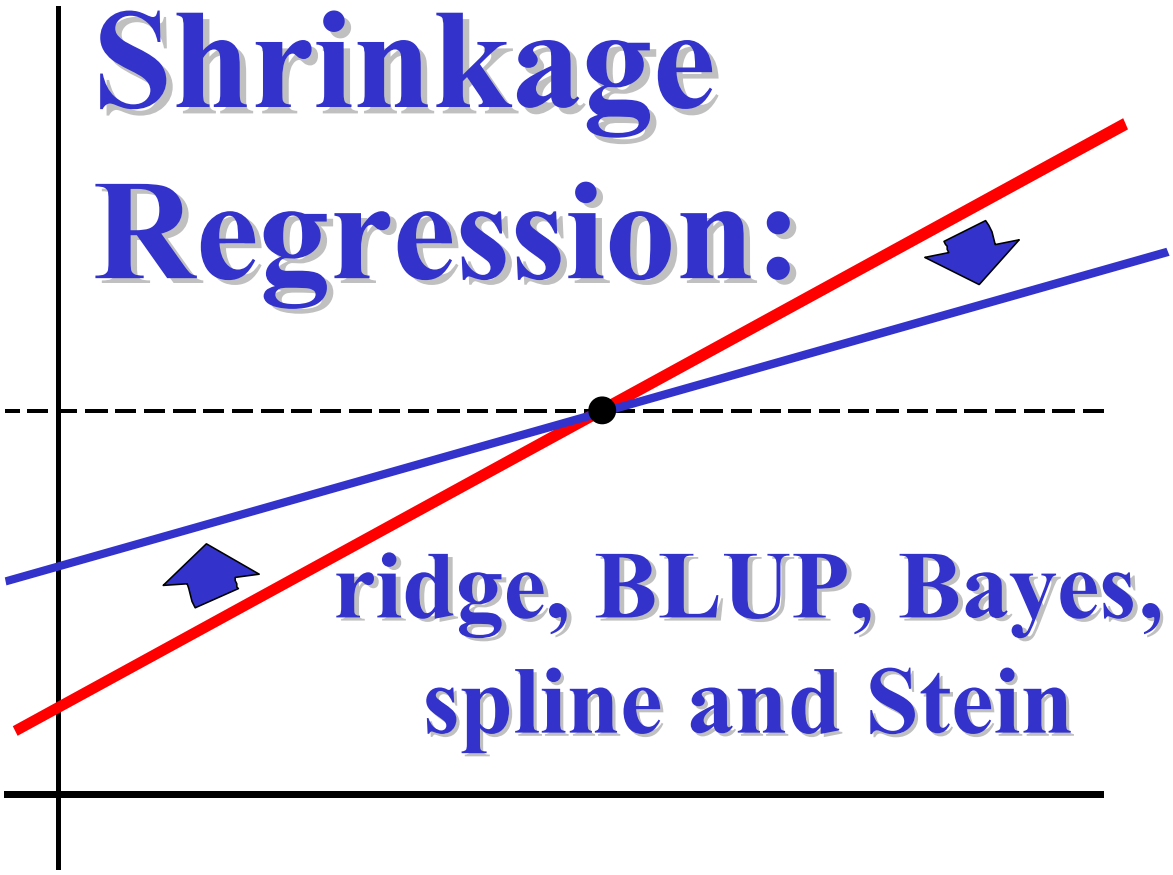


# Chapter 09: Computationally Intense Methods

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

# Chapter 9: Computationally Intense Methods (Errors-in-Variables, Resampling and Robustness)

The methodologies discussed in this chapter focus attention upon possible idiosyncrasies in observations of the independent X variables and/or the dependent Y variable of regression models. In addition to observational errors in the Y variable and possible ill-conditioning (high intercorrelations) among X variables that were considered in previous chapters, here we also consider data idiosyncrasies such as imprecise X values as well as "influential" observations (caused by outlying response values and/or high leverage regressor combinations.) Numerical algorithms commonly used to treat these sorts of idiosyncrasies tend to be computationally intense, and any data idiosyncracies detected by these methods can end up being either exploited or simply ignored!

A fitted regression solution is affected not only by (i) the available data but also by one's choices of (ii) statistical model, (iii) estimation methodology, and (iv) computational algorithms. Regression practitioners are well advised to explore the limitations imposed on accuracy and relevance of their fits by uncertainty stemming from all four of these sources. However, especially for the techniques discussed in this chapter, the final refrain from the Ballade of Multiple Regression, Corlett(1963), may well be particularly cogent:

"Your optimum only is bonum
*For the data you've fitted it to!"*

We start with two sections related to "errors-in-variables" models under which least-squares estimates are biased and inconsistent. Section §9.1 shows how random data un-rounding can suggest shrinkage to improve stability of estimates. But the maximum likelihood methods for multivariate normal "structural" models of section §9.2 suggest, instead, coefficient expansions to "correct-for-attenuation" resulting from uncertainty in predictor variables.

Next, we review iterative methods which, although traditionally applied to regressor variable subsetting or robust fitting, can also be used in shrinkage regression estimation. Section §9.3 discusses methods of cross-validation for choice and assessment of shrinkage. And "robust" methods that down-weight certain types of otherwise "influential" observations are described in section §9.4.

## 9.1 Regressor Perturbations After the Last Decimal Place

## and the Perturbation-Limit

This section discusses the main concepts introduced by Beaton, Rubin and Barone(1976) in their re-analysis of the infamous, ill-conditioned dataset of Longley(1967). Longley had used a six regressor model for U.S. economic results for the sixteen years from 1947 to 1962 to illustrate, rather dramatically, that the least-squares estimates from several software packages (apparently in widespread use at the time of his article) yielded estimates of poor numerical accuracy. As a direct result, being able to produce accurate results on the Longley “benchmark" is now widely considered some sort of minimal litmus-test for commercial statistical packages. After all, simple computational precautions like (internal) centering and scale-standardization of variables, calculating the singular-value-decomposition of X (rather than the eigen-decomposition of $X^TX$), and/or use of double precision suffice to “pass" this test.

I feel that the revelations of Beaton, Rubin and Barone(1976) were even more dramatic than those of Longley(1967). They showed (i) that small data perturbations (beyond the last decimal place reported in the Longley data) can yield even larger numerical changes in least-squares coefficient estimates than those from the computational algorithms Longley(1967) found to be least accurate. And they also showed (ii) that there is a well defined sense in which some of the estimates found to be least numerically accurate by Longley(1967) are actually more “reasonable" than the most accurate estimates! The following is a summary of the insights provided by Beaton, Rubin and Barone(1976) [B-R-B].

In a given dataset, such as the Longley(1967) benchmark, we may presume that the reported values are absolutely accurate as far as they go. In reality, the reported values have usually been rounded, possibly in some quite subjective way. For example, the Gross National Product (GNP) for 1947 of Longley(1967) was $X_2 = 234{,}289$ (or $234,289,000,000.00.) Rather than being precisely this value, the true GNP for 1947 may well have been almost any value between $234,288,500,000.00 and $234,289,499,999.99. Even a time-trend variable like Longley's $X_6 = \text{YEAR}$ is not a precise value in the sense that different years can contain different numbers of business days.

Now consider adding uniformly-distributed pseudo-random numbers on $[-0.5, +0.5)$, starting in the digit following the last published X digit. Small “errors" like these may seem almost trivial. After all, datasets generated in this way would be identical to the given data when rounded to the given number of digits. In this sense then, the un-rounded data are just as likely to be the exact data as are the given, rounded data.

Six of the 15 pairwise correlations among the six regressor variables in Longley's benchmark exceed $+0.98$. Because two of Longley's six X variables are reported with six significant figures and all have at least 3 significant figures, adding uniformily-distributed random deviates confined to the digits following the last reported decimal place assures that each set of perturbed regressor coordinates will remain intensely ill-conditioned and, thus, numerically unstable. And B-R-B found that the numerically accurate solution for the rounded (unperturbed) Longley data is “nowhere near the center of the distribution of a large number of presumably equally plausible [perturbed] solutions."

Consider the regression model $y = T \cdot \beta + \epsilon$ where (i) y is the $N \times 1$ vector of observations on the dependent variable, (ii) T is the $N \times p$ matix of true regressor values, (iii) $\beta$ is the $p \times 1$ vector of parameters to be estimated, and (iv) $\epsilon$ is a $N \times 1$ "catchall" vector for the unpredictable portion of y. The least-squares estimate of $\beta$ would then be $\widehat{\beta} = (T^T T)^{-1} T^T y$. Alternatively, letting X denote the observed, rounded version of T and writing $E = T - X$ for the difference between T and X, we have

$$\widehat{\beta} = [(X + E)^T (X + E)]^{-1} (X + E)^T y = [X^T X + E^T X + X^T E + E^T E]^{-1} (X^T y + E^T y).$$

Unfortunately, the numerical values of the T and E matrices are unknown. However, we may simulate the true least-squares estimates using a sequence of E matrices that represent statistically independent, uniformly-distributed data un-roundings. It then follows, for example, that $E(E) = 0$, $E(E^T X) = E(X^T E) = 0$ for fixed X, and $E(E^T y) = 0$ for fixed y. From similar expressions for fixed moments of order four or less, it follows that

$$E(\widehat{\beta}) = [X^T X/N + E(E^T E/N)]^{-1} X^T y/N, \qquad \{ 9.1 \}$$

where the expected value of $E^T E/N$ is the diagonal matrix of known variances of the regressor variable un-roundings. (Because the uniform distribution on $[-0.5, +0.5)$ has mean zero and variance $1/12 = 0.083\overline{3}$, un-rounding added following the d-th place after the decimal would have variance $0.083\overline{3} \times 10^{-2d}$.)

B-R-B call { 9.1 } the P-lim (perturbation limit) of their simulations and note that it is a Hoerl-Kennard(1970) ridge (shrinkage) estimator like that of our equation { 3.6 }, except that the diagonal elements of the $E(E^T E/N)$ matrix may not be all equal. Because $X^T X/N$ will have large off-diagonal elements (and at least some numerically small eigenvalues) when regressors are highly intercorrelated, the addition through $E(E^T E/N)$ of even relatively small (but positive) numerical values to the diagonal of $X^T X/N$ can create dramatic numerical changes in $E(\widehat{\beta})$ relative to the un-perturbed least-squares vector, $[X^T X/N]^{-1} X^T y/N$. In particular, these numerical changes can include even (i) changes in the numerical signs of some stabalized coefficients and (ii) changes of more than 5 digits in the second most significant figure of many coefficients.

In summary, then, Beaton, Rubin and Barone(1976) argue that the numerically most accurate least-squares solution corresponding to an ill-conditioned dataset can be extreme and implausible relative to the general distribution of solutions that can result from random data un-rounding. Furthermore, the (P-lim) "center" of this un-rounding distribution for coefficeints corresponds to a shrunken version of the un-perturbed least-squares coefficient estimates.

## 9.2 Multivariate Normal Errors-in-Variables Analyses

The presence of unknown measurement errors in the observed values of regressor variables leads to distortions in regression coefficient estimates; see, for example, Cochran(1968), Fuller(1987) and Gleser(1992). In the most simple case of a single (nonconstant) regressor

variable, $P = 1$, it is well known that the expected value of the resulting slope estimate is reduced in absolute value relative to the slope expected from regression on x-values free of measurement error; this effect is commonly called attenuation, as in Fuller and Hidiroglou(1978). Here, we outline some of the most obvious effects of measurement errors on coefficient estimates of multiple regression models and show that the resulting bias frequently corresponds to various forms of shrinkage of expected coefficient values. Correcting for this sort of bias in least-squares estimates motivates certain expansions of coefficient estimates, with corresponding inflations in their estimated variances.

Linear Errors-In-Variables Regression, EIVR, models (before centering) are of the form:

$$y = \tau + \epsilon \qquad \{ 9.2 \}$$

where

y is the $N \times 1$ vector of observations on the dependent variable,
$\epsilon$ is the $N \times 1$ vector of errors in the dependent variable,

and

$\tau = 1 \cdot \alpha + T \cdot \beta$ for $X = T + F$.

## 9.3 Cross-Validation, Bootstrapping, and Predictive Sample Reuse

### 9.3.1 Allen's PRESS Criterion.

### 9.3.2 A Rotation-Invariant Prediction Criterion.

## 9.4 Iterative Re-Weighting Methods

The results developed in Section §2.11, Analyses of Residuals, are sufficiently detailed to make them immediately applicable to most iterative robust-regression algorithms.

# References for Chapter Nine


Allen, D. M. (1974). "The relation between variable selection and data augmentation and a method for prediction." **Technometrics** 16, 125-127.

Andrews, D. F. (1974). "A robust method for multiple linear regression." **Technometrics** 16, 523-531.

Askin, R. G. and Montgomery, D. C. (1980). "Augmented robust estimators." **Technometrics** 22, 333-341.

Atkinson, A. C. (1986). "Masking unmasked." **Biometrika** 73, 533-541.

Bassett, G. and Koenker, R. (1978). "Asymptotic theory of least absolute error regression." **Journal of the American Statistical Association** 73, 618-622.

Beaton, A. E. and Tukey, J. W. (1974). "The fitting of power series, meaning polynomials, illustrated using band-spectroscopic data." **Technometrics** 16, 147-179.

Beaton, A. E., Rubin, D. B. and Barone, J. L. (1976). "The acceptability of regression solutions: another look at computational accuracy." **Journal of the American Statistical Association** 71, 158-168.

Berk, K. N. (1978). "Comparing subset selection procedures." **Technometrics** 20, 1-6.

Chambers, J. M. (1972). "Stabilizing linear regression against observational error in the independent variates." Bell Telephone Laboratories, Murray Hill, NJ.

Chambers, J. M. (1973). "Linear regression computations: some numerical and statistical aspects." **Proceedings of the 39th Session of the International Statistical Institute**, 45, 245-254.

Cochran, W. G. (1968). "Errors of measurement in statistics." **Technometrics** 10, 637-666.

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

Cook, R. D. and Weisberg, S. (1989). "Regression diagnostics with dynamic graphics." **Technometrics** 31, 277-291. [discussion 293-311.]

Corlett, T. (1963). "Ballade of multiple regression." **Applied Statistics** 12, 145.

Dallal, G. E., Rousseeuw, P. J., Leroy, A. M. and van Zomeren, B. C. (1991). **LMS: Least Median of Squares Regression**. FORTRAN Software for IBM-compatible Personal Computers.

Denby, L. and Mallows, C. L. (1977). "Two diagnostic displays for robust regression analysis." **Technometrics** 19, 1-13.

Fuller, W. A. and Hidiroglou, M. A. (1978). "Regression estimation after correcting for attenuation." **Journal of the American Statistical Association**, 73, 99-104.

Fuller, W. A. (1987). **Measurement Error Models**. New York, NY: John Wiley.

Gleser, L. J. (1992). "The importance of assessing measurement reliability in multivariate regression." **Journal of the American Statistical Association** 87, 696-707.

Gleser, L. J., Carroll, R. J. and Gallo, P. P. (1987). "The limiting distribution of least squares in an errors-in-variables model." **The Annals of Statistics** 15, 220-233.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984). "Location of several outliers in multiple-regression data using elemental sets." **Technometrics** 26, 197-208.

Henderson, H. V. and Velleman, P. (1981). "Building multiple regression models interactively." **Biometrics** 37, 391-411.

Hettmansperger, T. P. and McKean, J. W. (1977). "A robust alternative based on ranks to least squares in analyzing general linear models." **Technometrics** 19, 275-284.

Hettmansperger, T. P. and McKean, J. W. (1978). "Statistical inference based on ranks." **Psychometrika** 43, 69-79.

Hocking, R. R. (1972). "Criteria for selection of a subset regression: which one should be used?" **Technometrics**, 14, 967-970.

Hocking, R. R. (1976). "The analysis and selection of variables in linear regression." **Biometrics** 32, 1-49.

Holland, P. (1973). "Weighted ridge regression: combining ridge and robust regression methods." Working Paper #11, National Bureau of Economic Research, Cambridge, Massachusetts.

Kapenga, J. A. and McKean, J. W. (1988). "The vectorization of algorithms for R-estimates in linear models." **Proceedings of the 19th Symposium on the Interface: Computer Science and Statistics**, R. M. Heiberger, ed. 502-506.

Kapenga, J. A., McKean, J. W. and Vidmar, T. J. (1988). **RGLM: A Robust General Linear Model Package**, Version ASA-1.01. Kalamazoo, Michigan: Western Michigan University and The Upjohn Company.

Krasker, W. S. and Welsch, R. E. (1982). "Efficient bounded influence regression estimation." **Journal of the American Statistical Association** 77, 595-604.

Mason, R. L. and Gunst, R. F. (1985). "Outlier-induced collinearities." **Technometrics** 27, 401-407.

McKean, J. W. and Hettmansperger, T. P. (1976). "Tests of hypotheses of the general linear model based on ranks." **Communications in Statistics** A5, 693-709.

McKean, J. W. and Hettmansperger, T. P. (1980). "A robust analysis of the general linear model based on one-step R-estimates." **Biometrika** 65, 571-579.

McKean, J. W., Sheather, S. J. and Hettmansperger, T. P. (1976). "Regression diagnostics for rank-based methods." **Journal of the American Statistical Association** 85, 1018-1028.

Pariente, S. and Welsch, R. E. (1977). "Ridge and robust regression using parametric linear programming." Working Paper. Alfred P. Sloan School of Managemant, MIT, Cambridge, Massachusetts.

Pichard, R. R. and Berk, K. N. (1980). "Data splitting." **The American Statistician** 44, 140-147.

Pichard, R. R. and Cook, R. D. (1984). "Cross-validation of regression models." **Journal of the American Statistical Association** 79, 575-583.

Ramsay, J. O. (1977). "A comparative study of several robust estimates of slope, intercept, and scale in linear regression." **Journal of the American Statistical Association** 72, 608-615.

Roecker, E. B. (1991). "Prediction error and its estimation for subset-selected models." **Technometrics** 33, 459-468.

Rousseeuw, P. J. (1984). "Least median of squares regression." **Journal of the American Statistical Association** 79, 871-880.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). "Unmasking multivariate outliers and leverage points." **Journal of the American Statistical Association** 85, 633-639. [discussion 640-651.]

Walker, E. (1989). "Detection of collinearity-influential observations." **Communications in Statistics** A18, 1675-1690.

Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling plans in regression analysis." **The Annals of Statistics** 14, 1261-1295.

# Additional Reading for Chapter Nine


Afifi, A. A. and Azen, S. P. (1972). **Statistical Analysis: A Computer Oriented Approach.** New York: Academic Press.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). **Regression Diagnostics**. New York: John Wiley and Sons.

Chambers, J. M., Cleveland, W., Kleiner, B. and Tukey, P. (1983). **Graphical methods for data analysis**. Monterey, California: Wadsworth and Brooks-Cole.

Cook, R. D. and Weisberg, S. (1982). **Residuals and Influence in Regression**. New York: Chapman and Hall.

Daniel, C. and Wood, F. S. (1971). **Fitting Equations to Data: Computer Analysis of Multifactor Data for Scientists and Engineers**. New York: Wiley-Interscience.

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

Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. (1986). **Robust Statistics**. New York: John Wiley and Sons.

Huber, P. J. (1981). **Robust Statistics**. New York: John Wiley and Sons.

L'Ecuyer, P. (1988). "Efficient and portable combined random number generators." **Communications of the ACM** 31, 742-749,774.

Park, S. K. and Miller, K. W. (1988). "Random number generators: good ones are hard to find." **Communications of the ACM** 31, 1192-1201.

Press, W. H., Flannery, B. P., Teukolsky, S.A., and Vetterling, W. T. (1988). **Numerical Recipes in C: The Art of Scientific Computing**. [especially Chapter 7: Random Number Generation.] Massachusetts: Cambridge University Press. {Code Copyright 1985, 1987 by Numerical Recipes Software P. O. Box 243, Cambridge, MA 02238.}

Rousseeuw, P. J. and LeRoy, A. M. (1987). **Robust Regression and Outlier Detection**. New York: John Wiley and Sons.

Wichman, B. A. and Hill, I. D. (1982). "Algorithm AS 183: An efficient and portable pseudo-random number generator." **Applied Statistics** 31, 188-190.