Correction and Extension of

Wainer’s “Estimating Coefficients in Linear Models: It Don’t Make No Nevermind”

William M. Grove

University of Minnesota, Twin Cities

Draft of November 24, 2003: Do not cite or quote without permission
Abstract

In a citation classic, Wainer (1976) claimed that the expected loss in predictable variation due to using equal linear regression weights (applied to standardized predictors), instead of using optimal $\beta$ weights, is linearly increasing in $k$, the number of predictors. Wainer claimed to prove this while assuming that weights are independent and follow a uniform distribution on the interval (1/4, 3/4). Wainer erred twice: (1) he underestimated the loss by a factor of two (Laughlin, 1978), and (2) his posited weight distribution is infeasible for $k \geq 2$. When these errors are corrected, then for uniform (or symmetric beta distributed) optimal regression weights, one finds that the expected loss is nonincreasing as $k$ increases—the opposite of Wainer's result. I go on to examine the more general situation where regressors are correlated, and evaluate the expected loss by Monte Carlo integration. Only under highly implausible assumptions, about the distribution of the optimal regression weights, does one obtain expected loss notably greater than .025; this loss is nearly independent of $k$. 
Correction and Extension of

Wainer’s “Estimating Coefficients in Linear Models: It Don’t Make No Nevermind”

In a classic article, Wainer (1976) advanced two conclusions. First, he argued that the loss of predictable variation of a random variable, caused by using equal weights for regressors instead of optimal, possibly unequal weights, is likely to be small. He offered a putative proof of this, involving a second claim that the expected loss was at most $k/96$ (where $k$ is the number of regressors), though he did not show all steps of his proof.

Laughlin (1978) pointed out an error of algebra, and argued that the expected loss is really $k/48$, twice what Wainer claimed. The purpose of this note is to show that Wainer’s proof is further in error whenever $k \geq 2$. In fact, the relationship between loss and $k$ depends on the correlations among the predictors. Nonetheless, plausibility considerations do establish that the loss of predictability from using equal weights is indeed generally small.

To demonstrate these points, the article proceeds in four steps. First, Wainer’s proof is corrected (following Laughlin, 1978), to obtain the correct general expression for the expected loss, requiring the existence of two finite moments and independent predictors. In the second step, it is shown that Wainer's
postulated uniform distribution of optimal weights (adopted by Laughlin, 1978, in his correction of Wainer) is infeasible for any \( k \geq 2 \). Third, the expected loss is derived for a class of feasible weight (and predictor intercorrelation) distributions: the symmetric beta distribution family. Finally, the results of Wainer are extended to describe the situation obtained when predictors are correlated. Analytic and numerical evaluations of expected loss are tabled for non-negative predictor intercorrelations, and uniformly distributed weights. These give upper bounds on the expected loss when the weight distribution is more concentrated than the uniform, e.g., all unimodal distributions.

Pruzek and Frederick (1978) argued that, owing to the restrictiveness of assumptions made by Wainer, his conclusions misrepresent the comparative efficiencies of equal vs. least-squares regression weights over important parts of the parameter space. They also related the least-squares weights to the predictor intercorrelations and the predictor-criterion correlations via a useful principal components-type reparameterization. Many other authors have weighed in (pun intended) with data and opinions about the general position advocated by Wainer: both in favor of it (Bloch & Moses, 1988; Burt, 1950; Dawes, 1979; Dawes & Corrigan, 1974; Dorans & Drasgow, 1978; Einhorn & Hogarth, 1975; Gulliksen, 1950; Raju, Reyhan, Edwards, & Fleer, 1990; Richardson, 1941; Schmidt, Johnson,

However, none of these authors directly addressed the question Wainer asked and tried, however imperfectly, to answer: By how much do optimal (least squares) regression weights outperform equal weights, across a relevant parameter space? For example, Keren and Newman (1980) point out (correctly) that when predictors are negatively uncorrelated (rather than uncorrelated, as assumed by Wainer), as can occur when there are “suppressor variables” (Horst, 1941), least squares can greatly outperform equal weights. However, these authors did not give empirical data to help us understand how often we can expect to encounter suppressor variables in the real world; and they did not quantify the size of the aggregate loss due to using equal weights when predictors are negatively correlated, or the extent to which this loss would be vitiated by small losses when predictors are independent; let alone quantify relative performance when predictors are positively intercorrelated.

The question of aggregate loss may seem at first an uninteresting one. Why not compare them on a case by case basis, and eschew averaging over situations which are anywhere from very slightly to very greatly different? The answer is in
two parts. First, the expected comparative efficiencies of least squares vs. equal weights depend on population parameters, not sample estimates of them. A case-by-case approach is only partially enlightened by sample estimates of the relevant parameters—and the light will be dim when one has small samples, as one often does in psychology. It is useful in such situations to have a valid generalization, about the loss due to using equal weights, for a particular prediction domain. This generalization acts somewhat like a Bayesian prior probability, in serving as an anchor for sample-based adjustments to one’s opinion about whether to prefer equal or least squares weights. Second, a population average does not lose all its value when the population is known to be heterogeneous. Social statistics rely heavily on this fact—the mean (or median) income is a useful statistic even though we know it varies with the values of many factors such as education and age. If the average loss, due to using equal weights, is sufficiently small over a relevant part of the entire parameter space, then this justifies the conclusion that the loss is not very large over very much of the space.

This paper sticks with Wainer’s original question about the size of aggregate (expected) loss in predictive efficiency, due to using equal instead of optimal least squares weights in linear regression. It corrects the error of Wainer (1976) previously pointed out by Laughlin (1978) (and admitted by Wainer, 1978), but
goes on to correct Laughlin's own error in relying on a mathematically impossible assumption. It extends Wainer's result by considering relative predictive efficiency of the two types of weights, when predictors are positively intercorrelated. Moreover, there is a particular focus on the behavior of the expected loss, due to using equal weights, as the number of predictors increases.

Recapitulation of Wainer's Argument

Suppose that a random variable \( y \), assumed without loss of generality to have zero mean and unit variance, is to be predicted by the linear regression equation

\[
y = \hat{y} + e = X\beta + e
\]  

(1)

where items in boldface upper case are matrices, while boldface lower case indicates a vector; \( \hat{X} \) is the best linear unbiased estimator of \( X \); \( X \) is an \( n \times k \) matrix of regressors (assumed here to be of full column rank). We assume, without loss of generality, that elements of the columns of \( X \) are sampled from populations with zero means and unit variances. \( e \) is a vector of independent, identically distributed residuals, with mean \( \mu = 0 \) and variance \( \sigma^2 \), and \( \beta \) is optimal in the least squares sense.

It is well known that \( \beta = \Sigma^{-1} \rho \), where \( \Sigma \) is the matrix of correlations
among the regressors and $\rho$ is the vector of zero-order predictor-criterion Pearson correlations. If predictors are uncorrelated, the optimal weights are just $\beta = \rho$.

By contrast, one could also predict $y$ via weights $\beta_i = \beta_j, i, j = 1, \ldots, k$, that is, all equal weights. The linear model is

$$y = \tilde{y} + \epsilon = aX1 + \epsilon, \quad (2)$$

where $1$ is a vector of ones. The value of $a$ that minimizes the sum of squared errors of prediction, $\epsilon' \epsilon$, is given by the solution of

$$\frac{de' \epsilon}{da} = 0.$$ 

The desired weight is $a = 1' \rho / 1' \Sigma 1$. If predictors are uncorrelated, the optimal weight is $a = \text{Mean}(\beta) = \text{Mean}(\rho)$, where $\beta$ and $\rho$ are regarded not as constants but as vector-valued random variables; $\text{Mean}(X)$ denotes the mean of the random variable $X$. Using this optimal equal weight, the squared multiple correlation coefficient is $R_c^2 = a^2 1' \Sigma 1$.

Consider $E[e'e - \epsilon' \epsilon]$, where $E[\cdot]$ is the expectation operator. (The expectation is taken over some plausible distribution for the elements of $\beta$. We discuss candidate distributions below.) Wainer calls this the expected loss, here written
E[loss]. Expressing E[loss] as a function of $k$, one obtains

$$E[\text{loss}] = R_0^2 - R_e^2$$

$$= \beta' \Sigma \beta - a^2 \frac{\rho' \mathbf{1} \mathbf{1}' \rho}{\mathbf{1}' \Sigma \mathbf{1}}$$

$$= \rho' \left( \Sigma - \frac{J}{a^2 \mathbf{1}' \Sigma \mathbf{1}} \right) \rho$$

(3)

where $R_0^2$ is the squared multiple correlation coefficient when optimal $\beta$ weights are used, and $R_e^2$ is the corresponding value when equal weights are used. $J = \mathbf{1} \mathbf{1}'$ is a $k \times k$ unit matrix. Importantly, these formulas are nonparametric in that they hold for all feasible identical distributions of $\beta_i$ ($i = 1, \ldots, k$). The second formula is obtained by substituting $\rho' \Sigma \rho$ for $\beta' \Sigma \beta$. In the special case of independent identically distributed (i.i.d.) predictor weights, $E[\text{loss}] = k \text{Var}(\beta)$, where $\text{Var}(X)$ is the variance of the random variate $X$. This result follows from noting that $a$ is the mean of the $\beta_i$.

**Effect of Varying $\beta_i$ Distributions on Expected Loss**

The sampling distribution for $\beta_i$ posited by Wainer is a uniform distribution on the interval $(1/4, 3/4)$. The variance of any uniform distribution is $(l_u - l_l)^2/12$, where $l_u$ and $l_l$ are the upper and lower limits of the support. Hence, $E[\text{loss}] = k/48$. Wainer (1976) gave a different value, $k/96$; Laughlin (1978) corrected this error.
This result may chagrin some readers since it says that, for large $k$, the loss due to equal weighting can become sizeable. In personality, attitude, and ability assessments, numerous items are often added together with equal weights, albeit on the raw score scale. Because Wainer considered equal weights for standardized (not raw) items, Wainer’s result is not directly applicable in that situation. Nevertheless, there is no real loss of mathematical generality involved, because the raw item scores can always be rescaled to zero means and unit variances, whereupon Wainer’s (corrected) theorem quite properly applies.

However, Wainer’s proof is still further in error. This is because the posited distribution of the $\beta_i$ cannot be correct for $k \geq 2$. For independent predictors, the squared multiple correlation coefficient is given by $R_0^2 = \sum_{i=1}^{k} \beta_i^2$ ($i = 1, \ldots, k$), since the regressors are standardized. $R_0^2$ cannot exceed unity. However, if Wainer’s distribution holds for $k = 2$, it directly follows that

$$\Pr\{\beta_1 > 1/\sqrt{2}, \beta_2 > 1/\sqrt{2}\} = \Pr\{R_0^2 > 1\}$$

$$= \frac{17 - 12\sqrt{2}}{4},$$

which exceeds zero and is therefore impossible. Indeed, for general $k$ and independent regressor weights, any identical distribution for the $\beta_i$ ($i = 1, \ldots, k$) whatever, if that distribution has support beyond $k^{-1/2}$, is impossible.

The importance of this bound is that it is a monotone decreasing function of
$k$. Hence, as $k$ increases (enlarging $E[\text{loss}]$) the upper bound on $\text{Var} (\beta)$ decreases; these forces tend to offset one another. Whether the loss is increasing, decreasing, or neither in $k$ depends on the precise distribution of the independent $\beta_i$.

One superficially appealing amendment to Wainer’s distribution will not banish the problem. Specifically, one cannot assume that the distribution of the $\beta_i$, in terms of $k$, is uniform on $(1/4, 1/\sqrt{k})$, in order to prevent $\Pr \{R^2_{0} > 1 \} > 0$. When $k \geq 16$ this distribution has zero-width support.

Let us instead consider distributions on $(0, 1/\sqrt{k})$, since these are in the spirit of Wainer’s original idea and they are feasible for general $k$. In particular, consider the uniform distribution on $(0, 1/\sqrt{k})$. Then one has

$$E[\text{loss}] = k \text{Var} (\beta)$$

$$= k \left( \frac{1}{12} \right) \left( \frac{1}{\sqrt{k}} - 0 \right)^2$$

$$= \frac{1}{12},$$

independent of $k$, rather than linearly increasing in $k$, as Wainer (1976) indicated.

Distributions other than the uniform, if the $\beta$’s are to have uniform sign and $\Pr \{R^2_{0} > 1 \} = 0$, must also have support on an interval no wider than $(0, 1/\sqrt{k})$. Hence, many commonly encountered continuous distributions like the normal, $\chi^2$, extreme value, or gamma (Pearson Type III) cannot be used as models for
the $\beta$s without truncation. However, the coincidentally named beta distribution has support on a finite interval, and gives flexible distribution shapes through its parameters $p$ and $q$.

Beta distributions need not be symmetric. However, inasmuch as Wainer’s distribution is symmetric, in this spirit the standardized symmetric beta distribution (Harter, 1978) is considered, with parameters $p = q$, $\mu$, and $\sigma^2$. This has support on $\left( \mu - \sigma \sqrt{2p + 1}, \mu + \sigma \sqrt{2p + 1} \right)$. The density is given by

$$p_X(x) = \frac{\Gamma(2p)}{\sigma (\Gamma(p))^2 \left( 2 \sqrt{2p + 1} \right)^{2p-1}} \left( 2p + 1 - \left( \frac{x - \mu}{\sigma} \right) \right)^{p-1}.$$  

When $p$ is below 1, the density is dish-shaped, while for $p$ greater than 1 the shape is concave. When $p = 1$, one has Wainer’s distribution, the uniform (but with bounds depending on $k$) as a special case. For values of $p$ considerably exceeding 1 (e.g., $p = 10$), the distribution becomes bell-shaped.

If one scales this distribution so it has support on $\left( 0, k^{-1/2} \right)$, then $\mu = k^{-1/2}/2$ and $p = 1/(3k)$. One can regard $\sigma^2$ as a constant and $p$ a function of $\sigma^2$, or vice versa. In the former case, note that for any member of the standardized symmetric beta family having support on the interval in question, $\sigma^2$ must lie in the interval $(0, 1/k)$ exclusive. This is because, if $\sigma^2 \geq 1/k$, then $p \leq 0$.

If, on the other hand, one considers $p$ a distribution shape constant independent of $k$, then $\sigma^2 = 1/(k(2p + 1))$; in particular, when $p = 1$, $\sigma^2 = 1/(3k)$.
Hence, $\sigma^2$ is inversely proportional to $k$. In this situation and with i.i.d. $\beta$s, the upper bound on $\sigma^2$ as a function of $k$ implies that the upper bound on expected loss is $E[\text{loss}] > 1/3$, i.e., a constant in $k$. This is also contrary to Wainer’s conclusion that $E[\text{loss}]$ is linearly increasing in $k$.

The choice of one symmetric beta distribution over another, for modeling variation of $\beta$ coefficients, will only affect the factor $\text{Var}(\beta)$ in Equation (4). It will not influence the general behavior of $E[\text{loss}]$ as a function of $k$.

For the nonce, it is therefore useful to employ a conservative strategy, namely assuming that the beta distribution in question is the uniform on $(0, 1/\sqrt{k})$. Among symmetric beta distributions, the uniform is the least favorable $\beta_i$ distribution for the use of equal weights, in the sense that it yields the largest value of $E[\text{loss}]$.

Note, however, that this strategy has a serious limitation, which tends to favor equal weights, some would say in an unreasonable fashion. Restricting the upper limit of the $\beta_i$ (for any symmetric beta distribution) to less than one has the additional effect, ceteris paribus, of reducing the variance of the $\beta_i$. As Above it was shown that, for independent regressors, $E[\text{loss}]$ is linearly related to $\text{Var}(\beta)$. Hence, one could interpret the chosen strategy as “stacking the deck” in favor of equal weights.
To answer this criticism, in a subsequent section a distribution of the $\beta_i$ is considered that has support on all of $(0, 1)$. This is achieved by dropping the assumption that $\beta_i$ ($i = 1, \ldots, k$) are i.i.d. Instead, it assumed only that the $\beta_i$ (and the covariances among the predictors $\sigma_{ij}$, $i, j = 1, \ldots, k; i \neq j$, if these are not zero), are jointly distributed with marginal support intervals $(0, 1)$, taking on values such that $R_o \leq 1$. This is equivalent to assuming (for the simple uncorrelated-predictor case) that the distribution of $\beta_1$ has support on $(0, 1)$; that of $\beta_2$ on $(0, 1 - \beta_1^2)$; that of $\beta_3$ on $(0, 1 - \beta_1^2 - \beta_2^2)$; and so on for $\beta_4, \ldots, \beta_k$.

Expected Loss for Correlated Regressors

Thus far, like Wainer this author has only considered uncorrelated predictors which, in turn, yield independent $\beta$s. This assumption is not only mathematically restrictive but also empirically implausible. Multiple predictors in psychology and related disciplines generally show nonzero correlations. In particular, in the ability, psychopathology, personality (single-trait) and attitude (single attitude) domains, predictors typically possess positive manifold. That is, when scaled so that their zero-order correlations with the criterion are all positive, the intercorrelations between predictors are also positive.

This is not to say that important exceptions do not exist. For example, Garb (1985) obtained negative correlations among predictors of schizophrenia. This
likely occurred because most patients in his sample did not have schizophrenia. They were positive for several of the DSM criteria (e.g., the one specifying that symptoms are not attributable to organicity, and the one requiring that symptoms not be attributable to major affective disorder) but they were negative on criterion A of the DSM (i.e., they did not have specifically schizophrenia-like psychotic symptoms). In view of the relative paucity of such counterexamples, we will proceed on the assumption of positive manifold, or at any rate the lack of stable negative predictor intercorrelations.

As noted above, the equal regression weight \( a \) is given by \( a = 1^\prime \rho / \sqrt{1^\prime \Sigma 1} \), whether regressors are uncorrelated or have arbitrary intercorrelations. The loss in the general, correlated predictors case is, for fixed \( \Sigma \) and \( \rho 

\[
\text{loss} = \beta^\prime \Sigma \beta - a 1^\prime \Sigma a 1 \\
= \beta^\prime \Sigma \beta - \left( \frac{1^\prime \rho}{1^\prime \Sigma 1} \right)^2 1^\prime \Sigma 1 \\
= \beta^\prime \Sigma \beta - \frac{\beta^\prime \Sigma 1 1^\prime \Sigma \beta}{1^\prime \Sigma 1} \\
= (\beta^\prime - a 1^\prime) \Sigma (\beta^\prime + a 1^\prime)
\]

By definition, the expected value of the loss is

\[
E[\text{loss}] = \int_0^\theta \left( R_o^2 - R_c^2 \right) f(\theta) d\theta_1 \ldots d\theta_{k(k+1)} \tag{5}
\]
over the (multidimensional) region of integration, \( \theta \) is a vector of length \( m = k(k + 1)/2 \) comprising the parameters \( \beta_1, \ldots, \beta_k \) and \( \sigma_{21}, \ldots, \sigma_{k \ k-1} \). \( f(\theta) \) is an appropriate probability density, such as the beta distribution adopted above.

One can evaluate this integral, taking \( f(\theta) \) to be a multivariate beta p.d.f., over a \( m \)-dimensional hypercube delimited by

\[
0 < \beta_i < \theta_{\text{max}}, \quad i = 1, \ldots, k, \\
0 < \sigma_{ij} < \theta_{\text{max}}, \quad i, j = 1, \ldots, k, \quad i > j.
\]

Here \( \theta_{\text{max}} \) is the greatest real solution of a certain cubic,

\[
k(k - 1)\theta_{\text{max}}^3 + k\theta_{\text{max}}^2 - 1 = 0,
\]

resulting from equating \( R^2 \) to unity, under the assumption that all elements of \( \theta \) equal \( \theta_{\text{max}} \). In this situation, \( R^2 = \beta^\top \Sigma \beta \), where

\[
\beta = [\theta_{\text{max}} \ldots \theta_{\text{max}}] \quad \text{and} \\
\Sigma = \begin{pmatrix}
1 & \theta_{\text{max}} & \cdots & \theta_{\text{max}} \\
\theta_{\text{max}} & 1 & \cdots & \theta_{\text{max}} \\
\vdots & \vdots & \ddots & \vdots \\
\theta_{\text{max}} & \theta_{\text{max}} & \cdots & \theta_{\text{max}}
\end{pmatrix}
\]

yielding the cubic

\[
R^2 - 1 = k\theta_{\text{max}}^2 + k(k + 1)\theta_{\text{max}}^3 - 1 = 0. \quad (6)
\]
It might seem that the assumption relied on above, namely that the distribution has $k(k + 1)/2$-dimensional support, cannot be correct. After all, the rank of the predictor-cum-predictand covariance matrix is, at most, $k + 1$. However, this objection fails to consider two key facts. First, the parameter dimensionality in question is not that of the p.d.f. of the random variable $Y$, or of the multivariate p.d.f. of the least squares regression weights $\beta_i$, but rather it is the number of linearly independent parameters governing the p.d.f. of the statistic $R^2$, treated as a random variable. The distribution of $R^2$ depends nonlinearily on the covariances $\sigma_{ij}$, and likewise nonlinearly on the weights $\beta_i$. Moreover, the range of these covariances and weights has been restricted so that no combination of covariance and weight values is impossible. Hence, within these ranges and for this p.d.f., the distribution is on $m = k(k + 1)/2$ dimensions, instead of degenerating into a smaller dimensional support space.

Hence, it is quite proper to use Equation (6) to derive values for $\theta_{\max}$. Consider that beta distribution having $p = 1$ for all $m$ parameters (i.e., the multidimensional uniform). For this distribution, the solution of Eq. (6) is $\theta_{\max}$. Table 1 gives values of the solution, as a function of $k$, for selected values up to 30.

Table 1 about here

The factor $f(\theta)$ of the integrand in Eq. (5), being constant over the hypercube
for a uniform p.d.f. (and equaling $\theta_{\text{max}}^m$), can be factored out, leaving the integrals of $R_o^2$ and $R_e^2$. For small $k$, the integral of $R_o^2$ has a compact expression; e.g., for $k = 3$ it is simply

$$\int_0^\theta R_o^2 = \theta_{\text{max}}^8 + \frac{3}{4} \theta_{\text{max}}^9.$$

Alas, even for $k$ as small as 3, $R_e^2$ has a significantly more cumbersome integral:

$$\int_0^\theta R_e^2 = \frac{1}{1536}(\theta_{\text{max}}^5(1168\theta_{\text{max}}^3(1 + \theta_{\text{max}}) - 27 \log 3

+ (3 + 2\theta_{\text{max}})^2(9 + 4\theta_{\text{max}}(3 + 5\theta_{\text{max}})) \log 3 + 2\theta_{\text{max}} - (3 + 4\theta_{\text{max}})^2)

(9 + 8\theta_{\text{max}}(3 + 4\theta_{\text{max}})) \log 3 + 4\theta_{\text{max}} + 27(1 + 2\theta_{\text{max}})^4 \log 3 + 6\theta_{\text{max}}).$$

At approximately $k \geq 5$, one has to resort to numerical integration for this integral (see appendix for details); the results thereof appear in Table 2.

**Table 2 about here**

From Table 2 it is apparent that the loss is always close to about 2.5%, which is even less than the $k/48$ figure for independent $\beta$s. (The reasons why $E[\text{loss}]$ is so low for positively correlated $\beta$s are discussed below, in connection with the theorem of Wilks, 1948.) It is also clear from the table that $E[\text{loss}]$ is a nearly constant function of $k$, at least in the tabled range.

**How to Get a Result Different from Wainer’s (Corrected) Result**

The results thus far are, of course, not identical to Wainer’s flawed ones, or
to Laughlin’s flawed correction of them. Nevertheless, they be most reasonably interpreted as largely supporting Wainer’s broad conclusion: “It don’t make no nevermind” whether optimal or equal weights are used.

It is, however, readily possible to construct situations for which the expected loss is much greater than in Table 2. Indeed, some authors (e.g., Keren & Newman, 1978; Pruzek & Frederick, 1978) have opined that Wainer’s assumed conditions are generally unrealistic, arguing that different assumptions show greater fidelity to empirical problems, while also producing typically higher losses than those shown in Table 2.

To obtain such results disfavoring equal weights, one need only integrate the $R^2$ functions in Eq. (5), to obtain the expected loss over selected subregions of the unit hypercube

$$0 < \beta_i < 1, \quad i = 1, \ldots, k, \quad 0 < \sigma_{ij} < 1, \quad i, j = 1, \ldots, k, \quad i > j.$$ 

The integrand is considered undefined for every point in the unit hypercube where $\Sigma$ is not positive definite, or $\Sigma$ and $\beta$ are jointly such that $R_0^2 > 1$.

With such a non-smooth integrand, there is no tractable closed form solution for the integral $E[\text{loss}]$, even for small $k$. Hence, it is useful to employ numerical integration. Monte Carlo and quasi-Monte Carlo methods could be
used in Mathematica, but Fortran is much faster. Accuracy of the computed \( E \text{loss} \), given in Table 3, is assessed by evaluating the variance of the integrand. The table gives figures truncated so as to be accurate, \( \pm 1 \) in the last digit, with probability .9.

Table 3 about here

To restrict attention to those parts of the unit hypercube where \( \Sigma \) is p.d. and \( \beta \Sigma \beta \leq 1 \) is equivalent to supposing a sampling distribution where the elements of \( \Sigma \) and of \( \beta \) are not independent—quite different from what was assumed in obtaining Table 2 results. Indeed, negative dependence between the elements of \( \beta \) holds, because high values of \( \text{beta}_1 \), say, are only possible with low values of the other \( \beta \)s; otherwise \( R^2_0 > 1 \) would occur.

It is therefore easy to see that the difference in results between Table 2 and Table 3 occurs chiefly for the following reason. One gets negative correlations between elements of \( \beta \), when one “clips off” corners of the unit hypercube region of integration to enforce the requirement \( R^2_0 \leq 1 \). These negative correlations induce quite high variance in the \( \beta \)s, greatly increasing the expected loss.

To amplify through an example, consider the simplest case, where \( k = 2 \), and with either zero or positive predictor intercorrelations. As the bound on \( \beta_1 \) goes higher and higher (from the Table 1 value of \( \theta_{\text{max}} \approx .57 \), up toward 1),
the typical “randomly sampled” values of \( \beta_1 \) will also go up. As they do, then unless one manages simultaneously to lower co-occurring “randomly sampled” values of \( \beta_2 \) sufficiently, one will encounter, with nonzero probability, coefficient vectors \( [\beta_1, \beta_2] \)' that yield \( R^2_o > 1 \).

As pointed out above, such occurrences cannot be prevented through appropriate choice of bounds on the sampling distribution, if one requires (as the author does) that each of the \( \beta \)'s has the same sampling distribution, and assumes that parameters vary independently. Instead, one must assign parameter vectors that yield \( R^2_o > 1 \) (e.g., \([\beta_1 = .99, \beta_1 = .99]', \sigma_{12} = 0\) ) a joint probability of occurrence of zero. On the other hand, \([\beta_1 = .99, \beta_1 = .01]', \sigma_{12} = 0 \) and \([\beta_1 = .01, \beta_1 = .99]', \sigma_{12} = 0 \) are permitted. From these remarks, it should be clear that selection-induced negative correlations occur between the elements of \( \beta \); they occur more mildly for positively intercorrelated predictors, but they still occur. These negative correlations grow stronger when the selection is stronger, which occurs as \( k \to \infty \).

In the literature on this topic, the author has not seen credible reason offered to suppose that, in empirical problems, the distribution of \( \beta \)'s shows negative dependence. For identically distributed \( \beta \)'s showing zero dependence, the results in Table 2 apply; Table 3 is just a mathematical curiosity.
Relation of These Results to Wilks (1938)

Wilks (1938) derived the expected correlation between weighted sums of random variates, under mild distributional assumptions. Denote any two distinct, randomly selected linear combinations of predictors $X$ by $y_g = X\beta_g$ and $y_h = X\beta_h$ ($g, h = 1, \ldots, \infty$). That is, $\beta_g$ and $\beta_h$ are sampled from appropriate distributions, not necessarily the same distribution. Wilks’s central theorem is as follows:

$$
E\left[ \text{Corr} \left( y_g, y_j \right) \right] = 1 - O(k^{-1}) \quad (7)
$$

where $\text{Corr} \left( g, h \right)$ is the Pearson correlation between random vectors $g$ and $vech$, $X$ is as in Equation (1) and $O(k^{-1})$ denotes remaining terms of order $k^{-1}$ or less.

The assumptions used to prove the theorem include: (1) the weight vectors are random; (2) the two weight vectors are independent of the intercorrelations among the predictors $\sigma_{ij}$; and (3) as $k \to \infty$, the number of positive $\sigma_{ij}$ grows as $k^2$. Importantly, independence of the weight vectors $\beta_g$ and $\beta_h$ from each other is not required for the result. In the present problem, $\beta_g$ is randomly sampled from its distribution whereas $\beta_h$ is a linear function of the $\beta_g$, far from being independent of $\beta_g$. For the chief sampling distribution considered here (the uniform), the correlation between $\beta_g$ and $\beta_h$ is, on average, $1/2$, independent of $k$. 

On the other hand, as Wilks points out, “It is to be noted that unless the nume
of positive correlations among the x's is of order n^2 [our k^2], then it is not
possible to construct a linear score which [has expected correlation with other
random linear scores asymptotically equal to one]” (1938, p. 27). The speed of
gence of the expected correlation is governed by the average of the \( \sigma_{ij} \).
For this reason, the Wilks theorem does not apply in Wainer’s original problem,
since he supposed that the predictor intercorrelations were all zero.

Ree, Carretta, and Earls (1998) cited Wilks’s theorem to justify the con-
clusion that differentially weighted regressors will not substantially excel equal
weights in decision-theoretic performance. Also, Laughlin (1978) cites Wilks
(1938) in making the argument that if the average \( \sigma_{ij} \) rises, it becomes more and
more unlikely to have optimal beta weights arise from Wainer’s originally pos-
tulated sampling distribution, namely uniform on the interval \((1/4, 3/4)\). This is
because higher interpredictor correlations will yield lower optimal beta weights,
unless the predictor-criterion correlations compensatorily go down as the inter-
predictor correlations go up. However, the expected correlation as a function of
the mean predictor intercorrelation is not the main point of the theorem; instead,
it focuses on the expected correlation as \( k \to \infty \).

Of course, Wilks’s conclusion can still be true even when neither linear
combination (e.g., optimal vs. equal weights) is independent of the predictor intercorrelations; the theorem gives sufficient, not necessary, conditions (aside from the third condition listed above, which is necessary).

In fact, our Table 2 (and 3) results have no particular relation to Wilks’s theorem. The independence of $k$ we observe is unlike the inverse dependence proved by the Wilks theorem. Our findings are instead due to countervailing influences. This is most easily seen by considering Wainer’s original case of independent $\beta$s: (1) the expected loss, due to using equal weights, increases as $k$, due to Wainer’s original general theorem (as corrected by Laughlin, 1978); (2) there is an upper bound on the distribution of the $\beta$s, which decreases as $\sqrt{k}$. In the range of $k$ considered here, these factors countervail (almost) precisely, so that the expected loss ends up being approximately a constant function of $k$.

Limitations of These Findings

If neither independence or regressors, nor positive manifold, holds, then the results obtained here will not hold. In particular, if there is negative manifold (which can only obtain for small $k$, the value of expected loss may be quite different. In this connection, the concept of suppressor variables (Horst, 1941) should be considered. In fact, Pruzek and Frederick (1978), in criticizing Wainer (1976), prominently featured this possibility. Keren & Newman (1978) also cite
this possibility as a limitation on the general attractiveness of using equal weights.

However, stable (replicated) suppressor variables seem to be quite uncommon in psychometric work in psychology. A preliminary survey on PsycLit for the phrase “suppressor variable” in either title or abstract, from 1887 to August 2002, yielded 60 hits (citations available on request), of which 49 were original empirical studies while 11 were meta-analytic, commentaries, or statistical or methodological in nature. Of the empirical studies, four lacked sufficient PsycLit information to determine what the suppressor and criterion variables were. A further nine were experimental studies of human judges’ use of constructed-to-be-suppressor cue data, i.e., they were not demonstrations of real-world suppressors. This left 36 categorizable empirical studies, of which two replicated each other (i.e., $1/35 < 3\%$ of studies having the same criterion and suppressor variable).

There are, of course, a number of possible explanations for this apparent paucity. These include: failure to use the key phrase “suppressor variable” either in the title or abstract of an article, report or dissertation, even though the actual document does contain a suppressor variable finding (including possibly a replication); failure to attempt to replicate reported suppressors; and failure to report replications of previously published suppressors (i.e., the file drawer effect, but for a positive type of result). However, it cannot be argued that the
apparent rarity of replicated suppressor variables is not prima facie evidence for their actual rarity. Consider, for example, that a rather strict conjunction of requirements must be met for a stable (replicable) classical suppressor effect to exist: stable correlation(s) between predictor variable(s) and the criterion, stable correlation (essentially zero, or even negative) between suppressor variable and criterion, and stable correlation(s) between predictor variable(s) and the suppressor variable. Similarly stable, though differently patterned, correlations must exist for other types of suppressor variables. In view of this, it is plausible that the unreplicated suppressor relationships reported in this literature represent unreliable, and hence fundamentally uninteresting, results.

We appear to lack more rigorous and systematic surveys to document how frequently real-world, replicated (let alone practically useful) suppressor effects are found. Hence, it appears that one can only look to case studies and expert opinion to get beyond the above preliminary survey result. One such case, in fact, concerns what is probably the best-known attempt deliberately to construct a scale that would usefully act as a suppressor in diagnostic classification: the original MMPI K scale (McKinley, Hathaway, & Meehl, 1948). The degree of suppression is indicated by the coefficient of K used to produced K-corrected scale scores, with originally reported non-zero coefficients varying from 0.2 to 1.0
(Dahlstrom, Welsh, & Dahlstrom, 1972, p. 129). Subsequent research, however, strongly suggests that the optimal K-corrections are actually zero for each scale (Barthlow, Graham, Ben-Porath, Tellegen, & McNulty, 2002); that is, the suppressor effect has not replicated at all well.

As to expert opinion, R.M. Dawes (personal communication, 2002) and Meehl (personal communication, 2002) also failed to turn up any replicable suppressor variables in the individual differences (clinical, counseling, industrial-organizational, behavior genetics) domain. Therefore, it would seem that the suppressor variable is, if certainly not a unicorn, *rara avis*.

Of course, the above represent only plausibility considerations, failing short of proof. On the other hand, protestations that Wainer's conclusions do not hold because the assumptions relied on here, either for independent regressors or for positively intercorrelated ones, are likewise no more than plausibility arguments. The advantage of the present results is that, by considering not only independent regressors, which *are* rather implausible in most applied problems, but also regressors having positive manifold, we have extended the plausibility of Wainer's essential result to cover much more of the mathematically feasible parameter space.
References


Richardson, M.W. (1941). The combination of measures. In P. Horst (Ed.),


Appendix. Numerical Integration of E[loss]

In computing the integral underlying Table 2, one can readily obtain closed-form expressions for $R_o^2$, evaluable to arbitrary precision. Hence, the $E[\text{loss}]$ estimate accuracy is limited only by errors in integrating $R_o^2$. For $k = 5, 6, 7$, multidimensional adaptive quadrature (Genz & Malik, 1983) of $R_o^2$ was employed. This algorithm is most efficient, and has good error-bounding properties.

For $k = 8, 9$, quasi-Monte Carlo integration based on low-discrepancy sequences (Halton, 1970; Hammersley & Handscomb, 1964; Sloan & Woźniakowski, 1998) was used instead of adaptive quadrature. This is because Genz-Malik integration is inefficient in high dimensions, and because for moderate dimensionality of the integral, quasi-Monte Carlo integration is thought to be much more efficient than Monte Carlo integration. However, at or above $k = 10$, quasi-Monte Carlo results became non-monotonic in $k$, an incorrect result. Hence, ordinary Monte Carlo integration was used for $k = 10, 15, 25, 30$.

Mathematica version 4.1 (Wolfram, 1999) implements Genz-Malik adaptive quadrature, Monte Carlo, and low-discrepancy (quasi-Monte Carlo) integration methods. Therefore, this package was used to perform the computations underlying Table 2.

For Table 3, the problem is more complicated. Here, the integrand in the
expected value has a form akin to a step function; to be precise, the function is undefined whenever $\Sigma$ is not nonnegative definite or is indefinite, or whenever $R_0^2 > 1$. Although the region where the p.d.f. is defined is obviously convex, it so far from being a hyperrectangle. For what are at present unknown reasons, quasi-Monte Carlo integration on this integrand, over the restricted parameter-space region in question, yields negative values of the integral even though the integrand is everywhere nonnegative.

For this reason, Monte Carlo integration was used to obtain the figures in Table 3. However, Mathematica was not used for the upper reaches of Table 3, owing to its lack of speed. Instead, for $k \geq 7$ a Fortran Monte Carlo integration program was written. By induction on $k = 2, 3, \ldots$, it was established that the scaleup in computational effort, as a function of $k$, is as $\sqrt{2} n^{(k+2)/k}$. The program for computing entries in Table 3 (whether Mathematica or Fortran) proceeds by computing candidate sample points $(\Sigma, \beta)$ and then ignoring them if $\Sigma$ is not positive definite, or if $\beta' \Sigma \beta = R_0^2 > 1$. At $k = 10$, about 35,500 generated $(\Sigma, \beta)$ pairs are required to obtain one pair that satisfies $R_0^2 \leq 1$. Hence, we would expect that at $k = 20$, about 14 billion samples will be required per acceptable function evaluation. Since even $k = 10$ requires hours to run on a 750 MHz dual-CPU Pentium III workstation, it is unworkable to evaluate the integral for
$k$ much greater than 10. Even a supercomputer offers speedup, over what is obtainable with a workstation, of only about 1,000-fold, and is therefore not up to the task.

Therefore, the figures for $k = 15$ and above were obtained by evaluating the integral over a hypercube whose edges were half-integer multiples of the values of $\theta_{\text{max}}$ from Table 1. That is, the region of integration outside this hypercube was not sampled. This is precisely analogous to evaluating tail integrals of some probability distributions by cutting off the extreme tails, on the grounds that they have such small areas that they can safely be ignored. The rule was: when a multiple of $\theta_{\text{max}}$ was reached, say $c$, such that less than one acceptable sample per hour on the workstation was being obtained, the contribution to the integral, of the region “tail” region from $c$ to 1, was ignored. Just because samples in the region beyond this are so hard to obtain, they make scant contribution to the integral, even if the function evaluated at such a point is sizeable in magnitude. Since the value of $E[\text{loss}]$ cannot exceed 1, one need not worry that ignoring rare, but extremely large, function values will cause serious misestimation of the integral.

Mathematica and OpenMP Fortran 95 codes are available from the author on request.
Author Note

William M. Grove, Department of Psychology, University of Minnesota–Twin Cities campus.

The author would like to thank Paul E. Meehl, who pointed out the apparent contradiction between Wilks (1938) and Wainer (1976), study of which led to the present investigation. Thanks are also due Howard Garb, who pointed out significant exceptions to the author's generalization about psychopathology variables showing positive manifold.

Correspondence concerning this article should be addressed to William M. Grove, Department of Psychology, University of Minnesota–Twin Cities Campus, N218 Elliott Hall, 75 East River Road, Minneapolis, MN 55455-0344, or by e-mail at grove001umn.edu.
Table 1. Parameter Values Yielding $R^2 = 1$

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\theta_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.565</td>
</tr>
<tr>
<td>3</td>
<td>.426</td>
</tr>
<tr>
<td>4</td>
<td>.349</td>
</tr>
<tr>
<td>5</td>
<td>.301</td>
</tr>
<tr>
<td>6</td>
<td>.267</td>
</tr>
<tr>
<td>7</td>
<td>.242</td>
</tr>
<tr>
<td>8</td>
<td>.221</td>
</tr>
<tr>
<td>9</td>
<td>.205</td>
</tr>
<tr>
<td>10</td>
<td>.192</td>
</tr>
<tr>
<td>15</td>
<td>.147</td>
</tr>
<tr>
<td>20</td>
<td>.122</td>
</tr>
<tr>
<td>25</td>
<td>.106</td>
</tr>
<tr>
<td>30</td>
<td>.0944</td>
</tr>
</tbody>
</table>

Note. $k$ is the number of predictors, and $\theta_{\text{max}}$ a value such that, if all parameters have this value, $R_o^2 = 1$. 
Table 2. Expected Loss (in $R^2$) for Positively Correlated Predictors (Parameters Less Than $\theta_{\text{max}}$)

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\text{E[loss]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0190</td>
</tr>
<tr>
<td>3</td>
<td>0.0236</td>
</tr>
<tr>
<td>4</td>
<td>0.0250</td>
</tr>
<tr>
<td>5</td>
<td>0.0256</td>
</tr>
<tr>
<td>6</td>
<td>0.0257</td>
</tr>
<tr>
<td>7</td>
<td>0.0256</td>
</tr>
<tr>
<td>8</td>
<td>0.0254</td>
</tr>
<tr>
<td>9</td>
<td>0.0252</td>
</tr>
<tr>
<td>10</td>
<td>0.0249</td>
</tr>
<tr>
<td>15</td>
<td>0.0292</td>
</tr>
<tr>
<td>20</td>
<td>0.0319</td>
</tr>
<tr>
<td>25</td>
<td>0.0402</td>
</tr>
<tr>
<td>30</td>
<td>0.0206</td>
</tr>
</tbody>
</table>

Note.
Table 3. Expected Loss (in $R^2$) for Positively Correlated Predictors (Parameters Less Than 1)

<table>
<thead>
<tr>
<th>$k$</th>
<th>E[loss]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.0474</td>
</tr>
<tr>
<td>3</td>
<td>.0668</td>
</tr>
<tr>
<td>4</td>
<td>.0781</td>
</tr>
<tr>
<td>5</td>
<td>.0848</td>
</tr>
<tr>
<td>6</td>
<td>.0825</td>
</tr>
<tr>
<td>7</td>
<td>.0878</td>
</tr>
<tr>
<td>8</td>
<td>.0826</td>
</tr>
<tr>
<td>9</td>
<td>.0853</td>
</tr>
<tr>
<td>10</td>
<td>.082</td>
</tr>
<tr>
<td>15</td>
<td>.089</td>
</tr>
<tr>
<td>20</td>
<td>.085</td>
</tr>
<tr>
<td>25</td>
<td>.097</td>
</tr>
<tr>
<td>30</td>
<td>.096</td>
</tr>
</tbody>
</table>

Note.