Published in final edited form as:

Am Stat. 2015 ; 69(3): 254–255. doi:10.1080/00031305.2015.1056613.

Reply

Thaddeus Tarpey,

Professor in the Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435

R. Todd Ogden,

Professor in the Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032

Eva Petkova, and

Associate Professor in the Department of Child and Adolescent Psychiatry, New York University, New York, NY 10016-6023

Senior Researcher at the Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY 10962

Ronald Christensen*

Professor in the Department of Mathematics and of Statistics, University of New Mexico, Albuquerque, NM, 87131

It has been brought to our attention that there is some very interesting work related to this problem by Henmi and Eguchi (2004) and more recently by Cook, Forzani, and Rothman (2013) [CFR13] and by Meng and Xie (2014). Ding (2014) made the very interesting observation that in simple linear regression the least squares estimates are still maximum likelihood estimates even when the variance of the predictor is known. More immediately, Cook, Forzani and Rothman [CFR15] make some very interesting points in their letter.

One result of this is the bad news is that in Section 5 of our article (dealing with multiple regression) we gave the wrong covariance matrix for the “ad hoc” estimator

$$\tilde{\beta}^* \equiv \Psi_x^{-1} \hat{\Psi}_{xy},$$

wherein Ψ_x is the covariance matrix of x and $\hat{\Psi}_{xy}$ is the usual unbiased estimate of the covariance matrix between the predictors x and the dependent variable y . The correct covariance matrix is

$$\text{cov}(\tilde{\beta}^*) = \frac{\sigma_y^2 \Psi_x^{-1} + \beta \beta'}{n-1}.$$

*This work was supported by NIMH grant R01 MH099003.

The good news is that it remains clear that when $n \gg p$, the tendency is for the variability of $\tilde{\beta}$ to be larger than that of the least squares estimates that have

$$\text{cov}(\hat{\beta}) = \frac{\sigma_y^2 \Psi_x^{-1} - \beta' \Psi_x \beta \Psi_x^{-1}}{n - p - 2}.$$

In the numerators, $\text{cov}(\tilde{\beta}^*)$ adds a nonnegative definite matrix to $\sigma_y^2 \Psi_x^{-1}$ whereas $\text{cov}(\hat{\beta})$ subtracts a positive definite matrix from $\sigma_y^2 \Psi_x^{-1}$. The formula we gave is an upper bound for the actual covariance matrix of $\tilde{\beta}^*$,

$$\text{cov}(\tilde{\beta}^*) \leq \frac{\sigma_y^2 + \Psi_{xy}' \Psi_x^{-1} \Psi_{xy} \Psi_x^{-1}}{n - 1} \Psi_x^{-1}$$

in the sense that $A \succeq B$ means $A - B$ is nonnegative definite. The discrepancy arose because the form we used for the covariance matrix of a Wishart distribution is only valid when looking at linear combinations of the Wishart matrix that correspond to symmetric matrices. Details are contained in an online supplementary file.

We now respond to some of the points made in CFR15. We rewrite our formulae for the covariance matrices in terms of the squared multiple correlation coefficient, say,

$$\mathcal{R}_p^2 \equiv \frac{\beta' \Psi_x \beta}{\sigma_y^2}.$$

The covariance matrices for the least squares and ad hoc estimates become, for $n - p - 2 > 0$,

$$\text{cov}(\hat{\beta}) = \frac{\sigma_y^2 - \beta' \Psi_x \beta}{n - p - 2} \Psi_x^{-1} = \sigma_y^2 \Psi_x^{-1} \frac{1 - \mathcal{R}_p^2}{n - p - 2},$$

and

$$\sigma_y^2 \Psi_x^{-1} \frac{1}{n - 1} \leq \text{cov}(\tilde{\beta}^*) = \frac{\sigma_y^2 \Psi_x^{-1} + \beta \beta'}{n - 1} \leq \frac{\sigma_y^2 + \beta' \Psi_x \beta}{n - 1} \Psi_x^{-1} = \sigma_y^2 \Psi_x^{-1} \frac{1 + \mathcal{R}_p^2}{n - 1}$$

Clearly, the ad hoc estimator will be better whenever

$$\frac{1 + \mathcal{R}_p^2}{n - 1} < \frac{1 - \mathcal{R}_p^2}{n - p - 2} \quad (1)$$

or, equivalently, when

$$\frac{1 + \mathcal{R}_p^2}{1 - \mathcal{R}_p^2} \left(\frac{n-p-2}{n-1} \right) < 1.$$

The least squares estimate is the best estimate among those that are unbiased conditional on the x_s , but the ad hoc estimate is biased when conditioned on the x_s , it is only unbiased unconditionally. Some slightly unpleasant algebra (cf. online supplement) establishes that inequality (1) is equivalent to CFR15's inequality (4). CFR15 very interestingly point out realistic situations where the ad hoc estimate is better. The least squares estimate will be better whenever

$$\frac{1 - \mathcal{R}_p^2}{n-p-2} < \frac{1}{n-1}. \quad (2)$$

However CFR15, and in CFR13 even more so, focus on asymptotic results.

For simplicity, suppose the first moments are known and we want to estimate $\mu_y + (x_0 - \mu_x)' \beta$ where we think of x_0 as fixed. In this case

$$\text{var}[\mu_y + (x_0 - \mu_x)' \hat{\beta}] = \frac{1 - \mathcal{R}_p^2}{n-p-2} \sigma_y^2 (x_0 - \mu_x)' \Psi_x^{-1} (x_0 - \mu_x)$$

and

$$\frac{1}{n-1} \sigma_y^2 (x_0 - \mu_x)' \Psi_x^{-1} (x_0 - \mu_x) \leq \text{var}[\mu_y + (x_0 - \mu_x)' \tilde{\beta}^*] \leq \frac{1 + \mathcal{R}_p^2}{n-1} \sigma_y^2 (x_0 - \mu_x)' \Psi_x^{-1} (x_0 - \mu_x)$$

so that the ratio of the variances becomes

$$\frac{1}{1 - \mathcal{R}_p^2} \frac{n-p-2}{n-1} \leq \frac{\text{var}[\mu_y + (x_0 - \mu_x)' \tilde{\beta}^*]}{\text{var}[\mu_y + (x_0 - \mu_x)' \hat{\beta}]} \leq \frac{1 + \mathcal{R}_p^2}{1 - \mathcal{R}_p^2} \frac{n-p-2}{n-1}.$$

Again following CFR15, let $n \rightarrow \infty$ with $p/n \rightarrow f \in [0, 1)$. Using obvious notation,

$$\frac{1}{1 - \mathcal{R}_p^2} (1-f) \leq \lim_{n \rightarrow \infty} \frac{\text{var}[\mu_y + (x_0 - \mu_x)' \tilde{\beta}^*]}{\text{var}[\mu_y + (x_0 - \mu_x)' \hat{\beta}]} \leq \frac{1 + \mathcal{R}_\infty^2}{1 - \mathcal{R}_\infty^2} (1-f).$$

Note that $\mathcal{R}_p^2$ is increasing in p because the best (linear) predictor cannot get worse as you add more predictor variables. If either $f=0$ or $\mathcal{R}_p^2 \rightarrow 1$ regardless of the value of f , the lower bound is greater than 1 and least squares estimates are preferred. This is consistent with Ding (2014) showing that the simple linear regression least squares estimates are maximum

likelihood so that they should be asymptotically optimal, i.e., p is fixed. If, however, $\mathcal{R}_p^2$ converges to some number less than one, say $2/3$, then the ad hoc estimate is preferred whenever f is sufficiently large, in this case when $f > 0.8$. CFR15's inequality (4) indicates that n has to be small relative to p for the ad hoc estimate to work well. Our result suggests similarly that p has to be large relative to n for it to work well but only applies when CFR15's $h(p)$ is bounded. ($h(p)$ is bounded if and only if $\mathcal{R}_p^2$ is bounded below 1.) Another asymptotic model considered by CFR13 keeps $n - p > 0$ constant so that $f = 1$. In that case, the variance ratio limit is indeterminate when we also have $\mathcal{R}_p^2 \rightarrow 1$. But if $\mathcal{R}_p^2$ is bounded below 1, the ad hoc estimate dominates asymptotically.

Alas, this is only a first approximation. Estimating the first moments makes things considerably more complicated. Estimates of the regression surface (predictions) actually have the form $\bar{y} + (x_0 - \bar{x})'\hat{\beta}$ and $\bar{y} + (x_0 - \bar{x})'\tilde{\beta}$.

Under multivariate normality, $(s_y^2, S_{xy}, S_{xx}) \perp (\bar{y}, \bar{x})$ with $x_0 \sim N(\mu_x, \Psi_x)$ independent of all of them and assuming that an arbitrary unbiased estimate $\tilde{\beta}$ is a function of s_y^2 , S_{xy} , and S_{xx}

$$\begin{aligned} & \text{var}[\bar{y} + (x_0 - \bar{x})'\tilde{\beta}] \\ &= \text{var}[\bar{y}] + 2\text{cov}[\bar{y}, (x_0 - \bar{x})'\tilde{\beta}] + \text{var}[(x_0 - \bar{x})'\tilde{\beta}] \\ &= \text{var}[\bar{y}] - 2\text{cov}[\bar{y}, \bar{x}'\tilde{\beta}] + \text{var}[(x_0 - \bar{x})'\tilde{\beta}] \\ &= \frac{\sigma_y^2}{n} - \frac{2}{n}\Psi_{yx}E(\tilde{\beta}) + \text{var}[(x_0 - \bar{x})'\tilde{\beta}] \\ &= \frac{\sigma_y^2}{n} - \frac{2}{n}\beta'\Psi_x\beta + \text{var}[(x_0 - \bar{x})'\tilde{\beta}] \\ &= \frac{\sigma_y^2}{n}[1 - 2\mathcal{R}_p^2] + \text{var}[(x_0 - \bar{x})'\tilde{\beta}]. \end{aligned}$$

Only the last term is affected by the choice of unbiased estimate of the regression coefficients.

For the last term,

$$\begin{aligned} & \text{var}[(x_0 - \bar{x})'\tilde{\beta}] \\ &= \text{var}\{E[(x_0 - \bar{x})'\tilde{\beta} | \tilde{\beta}]\} + E\{\text{var}[(x_0 - \bar{x})'\tilde{\beta} | \tilde{\beta}]\} \\ &= \text{var}\{0\} + E\{\tilde{\beta}'\text{cov}(x_0 - \bar{x})\tilde{\beta}\} \\ &= E\left\{\tilde{\beta}'\left[\frac{n+1}{n}\Psi_x\right]\tilde{\beta}\right\} \\ &= \frac{n+1}{n}\text{tr}[\Psi_x\text{cov}(\tilde{\beta})] + \frac{n+1}{n}E(\tilde{\beta})'\Psi_xE(\tilde{\beta}) \\ &= \frac{n+1}{n}\text{tr}[\Psi_x\text{cov}(\tilde{\beta})] + \frac{n+1}{n}\beta'\Psi_x\beta. \end{aligned}$$

The computations above were inspired by CFR13's equation (24).

Using our forms for the covariance matrices

$$\frac{n+1}{n}\sigma_y^2 p \left[\frac{1}{n-1} \right] + \frac{(n+1)\sigma_y^2}{n} \mathcal{R}_p^2 \leq \text{var}[(x_0 - \bar{x})'\tilde{\beta}] \leq \frac{n+1}{n}\sigma_y^2 p \left[\frac{1 + \mathcal{R}_p^2}{n-1} \right] + \frac{(n+1)\sigma_y^2}{n} \mathcal{R}_p^2$$

and

$$\text{var}[(x_0 - \bar{x})'\hat{\beta}] = \frac{n+1}{n}\sigma_y^2 p \left[\frac{1 - \mathcal{R}_p^2}{n-p-2} \right] + \frac{(n+1)\sigma_y^2}{n} \mathcal{R}_p^2$$

so once again, inequality (1) implies that the ad hoc estimate is better and inequality (2) implies that the least squares estimate is better.

Finally, rather than simply comparing the least squares estimate and the ad hoc estimate, why not use them both? In simple linear regression consider an unbiased combination of the least squares and ad hoc estimates

$$\tilde{\beta} \equiv \alpha \hat{\beta} + (1 - \alpha) \tilde{\beta}.$$

We want to find the choice of α that minimizes the variance of this estimate. The individual variances are

$$\text{var}(\hat{\beta}) = \frac{\sigma_y^2}{\sigma_x^2} \frac{1 - \mathcal{R}_1^2}{n-3} \quad \text{and} \quad \text{var}(\tilde{\beta}) = \frac{\sigma_y^2}{\sigma_x^2} \frac{1 + \mathcal{R}_1^2}{n-3},$$

where $\mathcal{R}_1^2$ now reduces to the squared correlation between y and x . The variance of the combination is

$$\text{var}(\tilde{\beta}) = \alpha^2 \text{var}(\hat{\beta}) + (1 - \alpha)^2 \text{var}(\tilde{\beta}) + 2\alpha(1 - \alpha) \text{cov}(\hat{\beta}, \tilde{\beta}).$$

However, $\text{cov}(\hat{\beta}, \tilde{\beta}) = \text{cov}\{E[\hat{\beta}|X], E[\tilde{\beta}|X]\} + E\{\text{cov}(\hat{\beta}, \tilde{\beta}|X)\}$ where $\text{cov}\{\hat{\beta}, E[\tilde{\beta}|X]\} = 0$ and

$$\begin{aligned} E\{\text{cov}(\hat{\beta}, \tilde{\beta}|X)\} &= E\left\{\text{cov}\left((X_c' X_c)^{-1} X_c' y, \frac{1}{\sigma_x^2(n-1)} X_c' y \middle| X\right)\right\} \\ &= E\left\{\frac{\sigma^2}{\sigma_x^2(n-1)} (X_c' X_c)^{-1} X_c' X_c \middle| X\right\} \\ &= \frac{\sigma_y^2(1 - \mathcal{R}_1^2)}{\sigma_x^2(n-1)}. \end{aligned}$$

Thus,

$$\text{var}(\tilde{\beta}) = \alpha^2 \frac{\sigma_y^2}{\sigma_x^2} \frac{1 - \mathcal{R}_1^2}{n - 3} + (1 - \alpha)^2 \frac{\sigma_y^2}{\sigma_x^2} \frac{1 + \mathcal{R}_1^2}{n - 1} + 2\alpha(1 - \alpha) \frac{\sigma_y^2}{\sigma_x^2} \frac{1 - \mathcal{R}_1^2}{n - 1}.$$

Setting the derivative with respect to alpha equal to zero, we get

$$\alpha = \frac{\mathcal{R}_1^2(n - 3)}{(1 - \mathcal{R}_1^2) + \mathcal{R}_1^2(n - 3)}.$$

From this, the only time one would discard knowledge of σ_x^2 , i.e., set $\alpha = 1$ and use $\hat{\beta}$ is when $\mathcal{R}_1^2 = 1$. Similarly, one would only use $\tilde{\beta}$ when $\mathcal{R}_1^2 = 0$. Asymptotically, however, this converges to the least squares estimate, as it should based on Ding (2014).

References

- Cook R DennisForzani LilianaRothman Adam J. Prediction in abundant high-dimensional linear regression. *Electronic Journal of Statistics*. 2013; 7:30593088.
- Ding Ping. Comment on ‘‘Tarpey, T., Ogden, R. T., Petkova, E., and Christensen R. (2014), ‘A Paradoxical Result in Estimating Regression Coefficients,’ *The American Statistician*, 68, 271276.’. *The American Statistician*. 2014; 68:316.
- Henmi MasayukiEguchi Shinto. A Paradox Concerning Nuisance Parameters and Projected Estimating Functions. *Biometrika*. 2004; 91:929–941.
- Meng Xiao-LiXie Xianchao. I Got More Data, My Model is More Refined, but My Estimator is Getting Worse! Am I Just Dumb? *Econometric Reviews*. 2014; 33(14):218250.