BIAS IN MULTIPLE REGRESSION WITH EXAMPLE FROM FEED EFFICIENCY

D.L. Robinson

Animal Genetics and Breeding Unit, NSW Agriculture, NSW Agriculture and CRC for the Cattle and Beef Industry, University of New England, Armidale, NSW 2351

INTRODUCTION

The issue of bias in multiple regression is not new. However, researchers are often tempted to ignore the problem, hoping that any biases will have only a small effect on results. The aim of this paper is to draw attention to problems of bias and show that, in some cases, such as feed efficiency testing, biases can be important. Formulae to correct for bias can be derived from the least squares equations. In an example using data from beef cattle in northern New South Wales, Australia, phenotypic regression equations for residual feed efficiency differed from equations derived from genotypic regression and feed standards formulae. This paper investigates whether correcting for bias provides more consistent estimates.

THEORY

When fitting the standard least squares regression equation :

$$y = X\beta + \text{error}$$

the estimate $\hat{\beta}$ of β, is unbiased only if the elements of the matrix X (which represent factor levels or covariate values), are recorded without error. Not all covariates contain measurement error. For example, a covariate for the number of days since the start of an experiment should be error free. In other cases, such as when a covariate represents the length or size of an object, measurement errors are often relatively small, so biases are often ignored.

However, as will be demonstrated later, the effect of bias is not always negligible. For the case of simple univariate linear regression : y = constant + βx + error, the amount of bias is easily calculated. Suppose we cannot measure x exactly, but have an approximation, $s = x + \eta$, where η is an independent error, so we must fit : y = constant + $\beta_e s$ + error, then

$$\hat{\beta}_e = \text{cov}(y,s)/\text{var}(s) = \beta\, \text{var}(x)/(\text{var}(x) + \text{var}(\eta))$$

where var = variance, cov = covariance and the subscript e has been added to the regression coefficient to signify that the equation was fitted using an independent variable containing error. The estimated regression coefficient, $\hat{\beta}_e$ is biased downward by the factor 1 / (1 + var(η) / var(x)). Thus if var(η) is of similar magnitude to var(x), the estimated coefficient will be approximately half the true value.

The situation is more complicated when two or more independent variables are involved. For example, suppose $y = \alpha a + \beta b$ + error, and that measurement errors in a are negligible. Also suppose we cannot measure b, only $B = b + \delta$ where δ represents an independent error term with variance D. It is therefore necessary to fit : $y = \alpha_e a + \beta_e B$ + error. The above example assumes, for simplicity, that a, b and the error terms have mean zero, so it is not necessary to fit intercepts. Solving the least squares equations yields the estimates :

$$\hat{\alpha}_e = \alpha + \beta D c_{ab} / (v_a (v_b + D) - c_{ab}^2)$$
$$\hat{\beta}_e = \beta (1 - v_a D / (v_a (v_b + D) - c_{ab}^2))$$
where v_a = var(a), v_b = var(b) and c_{ab} = cov(a,b) (1)

Thus, as in the univariate case, there is downward bias in the estimated partial regression coefficient for the term containing significant error. In contrast, if the two covariates, *a* and *b*, are positively correlated and β is positive, there is a corresponding upward bias in the estimated partial regression coefficient, $\hat{\alpha}_e$.

The amount of bias is illustrated in table 1 for a hypothetical example where var(a) = var(b)= β = α = 1.0, the covariance, c_{ab}, between *a* and *b* ranges from 0.1 to 0.9 and D, the error variation in B, ranges from D = 0.5 (error in *B* equal to half the variation of the 'signal', var(b)) to D = 3.0 (3 times as much error or 'noise' as signal). Table 1 shows that the estimated partial regression coefficient for *B* has greater bias than that for *a* and that the bias increases when *a* and *b* are highly correlated. When c_{ab} = 0.9, the estimate, $\hat{\beta}_e$ is only 28 % of its true value if D = 0.5, and only 6 % of its true value if D = 3.0 (table 1). The estimated partial regression coefficient for *a* is biased upward, being 35 % higher than its true value when c_{ab} = 0.7, D = 0.5 and 85 % higher if c_{ab} = 0.9, D = 3.0. Thus if covariates are known to contain error, it will often be useful to consider how the amount of error may affect estimated partial regression coefficients and whether this will influence the analysis of data or interpretation of results.

Table 1. Values of $\hat{\alpha}_e$ and $\hat{\beta}_e$ when var(a) = var(b) = β = α = 1.0, c_{ab} ranges from 0.1 to 0.9 and D = 0.5, 1.0 or 3.0

c_{ab}	D	$\hat{\alpha}_e$	$\hat{\beta}_e$	c_{ab}	D	$\hat{\alpha}_e$	$\hat{\beta}_e$	c_{ab}	D	$\hat{\alpha}_e$	$\hat{\beta}_e$
0.1	0.5	1.03	0.66	0.1	1	1.05	0.50	0.1	3	1.08	0.25
0.3	0.5	1.11	0.65	0.3	1	1.16	0.48	0.3	3	1.23	0.23
0.5	0.5	1.20	0.60	0.5	1	1.29	0.43	0.5	3	1.40	0.20
0.7	0.5	1.35	0.50	0.7	1	1.46	0.34	0.7	3	1.60	0.15
0.9	0.5	1.65	0.28	0.9	1	1.76	0.16	0.9	3	1.85	0.06

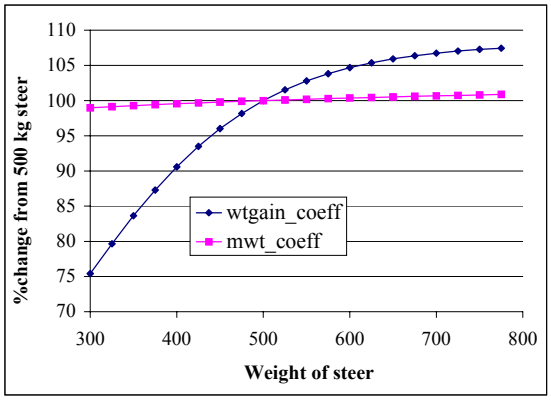
EXAMPLE USING FEED EFFICIENCY OF BEEF CATTLE

A trait often used for the calculation of feed efficiency is residual feed intake (RFI). RFI is defined as the amount of feed eaten by an animal less what would be expected from the growth of the animal and its body weight (used as an indicator of maintenance requirements). Kennedy *et al*. (1993) suggested that RFI could be based on genotypic regression, though it is more commonly calculated from feeding standards formulae (e.g. Arthur *et al.*, 2001), or as the error term when fitting the equation :

FI = constant + α*mwt + β*wtgain + error (i.e. RFI) (2)

where FI = feed intake(kg/day) ; **mwt** is the metabolic weight of the animal, expected to be proportional to its maintenance requirements ; mwt is calculated as mean(weight$^{0.73}$), or sometimes midweight$^{0.73}$ where midweight is the animal's weight at the midpoint of the test period, and **wtgain** is the estimated weight gain of the animal over the period for which feed

intake was measured. In general, mwt should contain very little error, especially if it is estimated from the mean of several measurements. In contrast, the accuracy of wtgain is dependent on the time interval over which weight gain is estimated (see e.g. Robinson, 2002).



Feed standards formulae. SCA (1990) provided formulae for feed intake of cattle by weight, weight gain, sex and breed. The published equations used weight$^{0.75}$ but have been converted to weight$^{0.73}$ for compatability with Equation 2. The figure (left) shows the change in SCA coefficients for a *Bos taurus* steer as animals become heavier. The coefficient for weight$^{0.73}$ changes very little. However, the SCA coefficient for wtgain indicates that a 300 kg steer should require only 75 % of the feed needed for a 500 kg steer to gain 1 kg of weight. In contrast, a 600 kg steer should require 5 % more feed to gain 1 kg. The change reflects the anticipated increase in the proportion of fat deposited per kg of weight gain as animals approach maturity.

Genotypic regression. Kennedy *et al.* (1993) suggested defining RFI as feed intake less what was predicted from genotypic regression of feed intake on production. In cases where phenotypic measurements such as weight gain are subject to considerable measurement error, genotypic regression may provide additional insight into the true relationship. Let G be the genetic (co)variance matrix of the two production traits (mwt and weight gain) and c be the vector of genetic covariances of feed intake with production, then the genotypic regression coefficients β_{gen} are given by : $\beta_{gen} = G^{-1}c$. Robinson and Oddy (2002) reported estimated genetic parameters for feed intake, mwt and feedlot weight gain (FLgain) from 1481 beef cattle feedlot-finished for the domestic (400 kg liveweight at slaughter), Korean (520 kg) and Japanese (steers only ; 600kg) markets. Solving for β_{gen} results in the equation : $RFI_{gen} = FI_{gen} - 0.044mwt_{gen} - 4.998FLgain_{gen}$ where RFI_{gen} is RFI calculated by genotypic regression and the subscript gen has been added to indicate applicability to genetic values.

Phenotypic regression. Robinson and Oddy (2002) also presented results for phenotypic regression of feed intake on mwt and FLgain for the same 1481 cattle using a model including fixed terms for mwt, FLgain, age and the overall mean, plus random terms for market, test group, breed type (tropical or temperate) and breed of cattle, birth herd, animal effects (including pedigree) and previous nutritional status. Because the relationship of feed intake with weight gain differed according to age and destination market (domestic, Korean or Japanese), an overall fixed slope was fitted for FLgain with random regression terms for the interaction with market. Random interactions were also fitted for mwt with market, sex and test group, but they had very little estimated variation and so minimal impact on the results.
FLgain was used, as described above, by Robinson and Oddy (2002) because weight gain in the automatic feeder (AF) pens (AFgain), which would normally be used to calculate RFI (see Equation 2), was considered too unreliable. The low accuracy of AFgain was due to the relatively short time cattle spent in the AF pens (53, 58 and 79 days respectively for domestic, Korean and Japanese markets). FLgain was estimated by fitting a fixed linear regression

model : weight = animal + animal.day + date of weighing. AFgain was estimated from a similar model, but also adjusting for feed intake on the day of weighing as an indicator of gut fill. Estimated error variances of FLgain and AFgain (calculated from the standard errors of regression coefficients in the linear regression models) were 0.0065 and 0.0396 $(kg/day)^2$ respectively.
The estimated phenotypic (co)variance matrix for mwt and FLgain was (32.4, 0.77, 0.056) (Robinson and Oddy, 2002). This implies that, for the terms defined in Equations 1, $v_a = 32.4$, $c_{ab} = 0.77$, $v_b = 0.0495$ and D = 0.0065. It was assumed that the same parameters applied to AFgain, except that D = 0.0396.

Effect of correcting for bias. Results from genotypic regression, the SCA equations and from phenotypic regression fitting FLgain and AFgain are shown in table 2, along with the conversion from phenotypic to 'unbiased' regression using Equations 1. Before conversion, results for FLgain and AFgain differed substantially, e.g. the coefficient for mwt was 0.115 using FLgain, compared to 0.153 fitting AFgain. Conversion more than doubled the estimated coefficients for AFgain. Although there is some variability (due to the many factors influencing the data), the converted estimates for weight gain appear to be generally consistent both with the SCA equations and estimates from genotypic regression. However, the coefficient for mwt was still substantially higher than the SCA and genotypic regression estimates ; but this cannot be explained by measurement errors in estimated weight gain of each animal.

Table 2. Estimates of the relationship between feed intake, mwt and weight gain

	$\hat{\alpha}_e$	$^1\hat{\beta}_{eD}$	$^1\hat{\beta}_{eK}$	$^1\hat{\beta}_{eJ}$	$\hat{\alpha}$	$^1\hat{\beta}_D$	$^1\hat{\beta}_K$	$^1\hat{\beta}_J$
Genotypic Regression					0.044	5.0[2]	5.0[2]	5.0[2]
SCA Equations					0.054	4.2	4.6	4.7
	Regression coefficients as fitted				Conversion (Equations 1)			
Feedlot wt gain	0.115	2.8	3.0	4.0	0.098	3.4	3.7	4.9
1) AFgain[3] (all data)	0.156	1.8	2.1	1.4	0.102	4.2	4.8	3.3
2) AFgain[3] (subset)	0.150	1.9	1.7	2.1	0.093	4.3	3.8	4.8
Mean of 1) and 2)	0.153	1.9	1.9	1.8	0.097	4.2	4.3	4.0

[1] Subscripts D, K and J refer to Domestic, Korean and Japanese markets. [2] Genetic parameters were derived from all markets, so genotypic regression represents an average for all markets. [3] The regression model (see paragraph with heading : 'Phenotypic regression') was fitted to 1) all data except 3 test groups for which weight records in the AF pens spanned less than 40 days ; 2) the last 27 of the 36 test groups for which feed intake measurements were available on the day of weighing until the time animals were weighed. All results (except genotypic regression) are for temperate breed steers.

REFERENCES

Arthur, P.F., Renand, G. and Krauss, D. (2001) *Livest. Prod. Sci*. **68** : 131-139.

Kennedy, B.W., van der Werf, J.H.J. and Meuwissen, T.H.E. (1993) *J. Anim. Sci.* **71** : 3239-3250.

Robinson, D.L. (2002) *(to be submitted)*.

Robinson, D.L. and Oddy, V.H. (2002) *(to be submitted)*.

SCA (1990) Standing Committee on Agriculture - Ruminants Subcommittee. Feeding Standards for Australian Livestock. CSIRO, Australia, 1990.