Misuse of coefficient of determination for empirical validation of models

M.J. McPhee and B.J. Walmsley

NSW Department of Primary Industries, Beef Industry Centre, University of New England, Armidale, NSW 2351, Australia
Email: malcolm.mcphee@dpi.nsw.gov.au

Abstract: The objective of this study was to demonstrate that the regression coefficient of determination (r^2) is misused when validating models. The deviance measures of mean square error of prediction (MSEP), the decomposition of MSEP (bias, slope, and random components), modelling efficiency, t-tests of bias and slope and the r^2 are all reported for 3 datasets.

On-farm data to evaluate BeefSpecs, a fat deposition model that predicts final P8 fat (mm) to assist producers meet market specifications, is the primary dataset used in this study [dataset 1 (n = 80)]. Datasets 2 (n = 80) and 3 (n = 80) were created from dataset 1. Three millimetres were added to the observed values of dataset 1 to create dataset 2 and for dataset 3 the observed and predicted values were rotated 90 degrees. For datasets 1, 2, and 3 respectively, the mean bias was 0.06, 3.06, and 0.06 mm, root-MSEP (RMSEP) was 1.72, 3.51 and 3.68 mm, bias was 0.1, 76, and 0%, slope was 6, 1.4, and 79%, random component was 94, 23, and 21%, modelling efficiency (MEF) was 0.39, -1.55, and -1.80 and the r^2 was 0.43, 0.43, and 0.43 and the percentage of data points within ± 1.5 mm of the control limits was 65, 14, and 31%.

This study strongly emphasizes that model validation should be conducted with the standard reporting of summary statistics of min, max, mean, standard deviation, mean bias, RMSEP followed by the decomposition of MSEP (bias, slope, and random components) along with the graphical display of observed vs predicted (Figure 1) and deviations with upper and lower control limits. The MEF is also recommended as an appropriate assessment of model evaluation in preference to r^2.

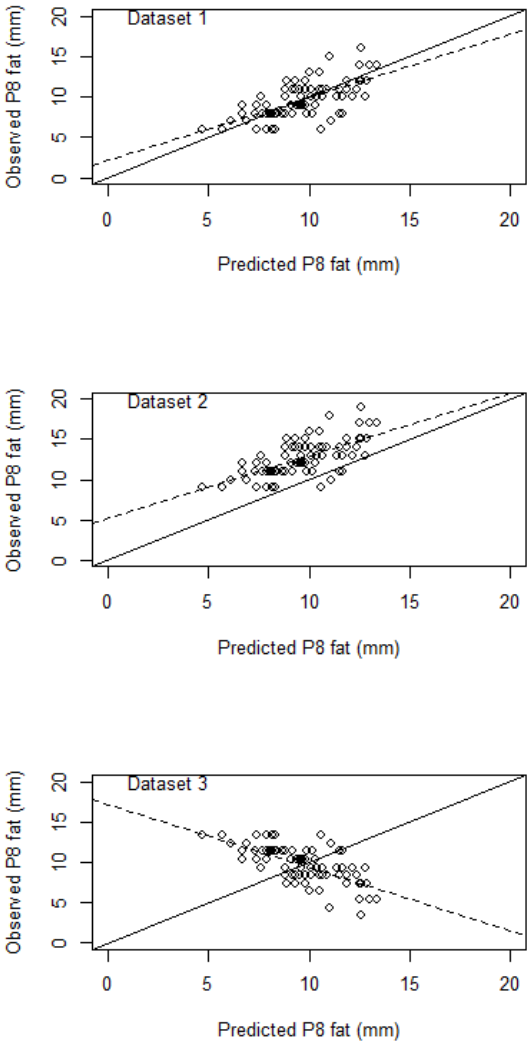

Figure 1. Datasets of the observed versus predicted P8 fat (mm) illustrating scenarios to demonstrate the misuse of regression for empirical validation of models. Dashed line is line of best fit.

Keywords: *Deviance measures, modelling efficiency, bias, slope, random*

1. INTRODUCTION

Several publications (Loague and Green, 1991; Mayer and Butler, 1993; Mitchell, 1997; Mitchell and Sheehy, 1997; Tedeschi, 2006; Pineiro et al. 2008) have been written on the statistical validation of models. Arguments against using regression for empirical validations (Mitchell, 1997) and graphical methods for evaluating models (Loague and Green, 1991; Mitchell, 1997) have been discussed. The review by Tedeschi (2006) covers a broad range of statistical techniques for model evaluation. This paper supports the argument by Mitchell (1997) and Mitchell and Sheehy (1997) that research scientists generally misuse regression for empirical validation of models. A subset of statistical techniques reported in a review by Tedeschi (2006) is presented in this paper along with an emphasis on the misuse of the coefficient of determination (r^2). The principal of reporting r^2 in publications is strongly held among research scientists and reviewers. When validating (challenging or assessing) the adequacy of a model, the reporting of r^2 is miss interpreted. This study uses real data from a BeefSpecs evaluation study (McPhee et al., 2014) to illustrate the misuse of the r^2 statistic for empirical validation of models.

2. THE STATISTICAL MODEL AND DEVIANCE MEASURES

The statistical model and several deviance measures used by research scientists to evaluate the accuracy and precision of models predicted from simulation and empirical models versus the observed value are reported in this section. Mathematical notation used in this paper is described in Table 1.

Table 1. Mathematical notation and description

Notations	Description
Deviation	The difference between observed and model-predicted values $(Y_i - f(X_1,...,X_p)_i)$
$f(X_1,...,X_p)_i$	The ith model-predicted (or simulated) value
$\overline{f}(X_1,...,X_p)_i$	Mean of model-predicted (or simulated) values
Y_i	The ith observed or measured value
$\overline{Y}$	Mean of the observed (or measured) values
$\hat{Y}$	The ith linear value of the evaluation regression
MSEP	Mean square error of prediction
RMSEP	Root mean square error of prediction
Bias	MSEP decomposed into error due to overall bias of prediction
Slope	MSEP decomposed into error due to deviation of the regression slope from unity
Random	MSEP decomposed into error due to the random variation
MEF	Modelling efficiency
r	Correlation coefficient
r^2	Coefficient of determination
SSR	Sums of squares of regression about the fitted line
SSE	Sums of squares of the error
SSTO	Sums of squares total

2.1. Statistical model

The linear regression (1) is commonly used to evaluate a model

$$Y_i = \beta_o + \beta_1 \times f(X_1,...,X_p)_i + \varepsilon_i \tag{1}$$

where β_o and β_1 are the regression parameters for the intercept and slope, respectively and ε_i is the ith random error assumed to be from a single population that is independent and normally distributed ~ $\mathrm{N}(0,\sigma^2)$.

2.2. Deviance measures

The MSEP (2) is the most common deviance measure. It compares the observed values versus the model-predicted values. The square root of the MSEP (RMSEP) is generally reported rather than the MSEP.

$$MSEP = \frac{\sum_{i=1}^{n}(Y_i - f(X_1,...,X_p)_i)^2}{n} \tag{2}$$

The decomposition of MSEP was first introduced by Theil (1966) and outlined with additional explanation by Bibby and Toutenburg (1977); the breakdown is expressed as errors in central tendency, errors due to regression and errors due to disturbances that sum to the MSEP i.e., MSEP = Bias (3) + Slope (4) + Random (5). Both the slope and random components represent the sample variance of predicted and observed values. The bias, slope, and random components are generally reported as percentages of the total MSEP.

$$Bias = (\overline{f}(X_1,...,X_p)_i - \overline{Y})^2 \tag{3}$$

$$Slope = \frac{\sum_{i=1}^{n}(f(X_1,...,X_p)_i - \overline{f}(X_1,...,X_p)_i)^2}{n} \times (1-\beta_1)^2 \tag{4}$$

$$Random = (1-r^2) \times \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n} \tag{5}$$

The MEF (6) described by Loague and Green (1991) and reported by Mayer and Butler (1993) is a dimensionless statistic which directly relates model predictions to observed data.

$$MEF = 1 - \frac{\sum_{i=1}^{n}(Y_i - f(X_1,...,X_p)_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{6}$$

The coefficient of determination (r^2) (7) is interpreted as the proportion of variation explained by the fitted line.

$$r^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{7}$$

and (7) can be rewritten as (8) because SSTO = SSR + SSE

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{8}$$

2.3. Quality control

Statistical quality control (Montgomery, 1991) techniques of upper and lower control limits where the control is set at 0 are employed here. The residuals (observed minus predicted) are reported where the associated error of upper and lower limits are set at 1.5 mm, which is the accepted proficiency level for accreditation to

register ultrasound scanners (Upton et al., 1999) for P8 fat depth measurements to enter the national beef genetic improvement scheme (Graser et al., 2005).

2.4. Statistical analysis

Three datasets (Table 2) were used to illustrate the miss use of r^2. Model evaluation of observed versus predicted P8 fat was conducted using a customized procedure in the R statistical package (R Development Core Team, 2009). The statistical significance of each mean bias was evaluated using a paired *t*-test of the mean of the differences between the observed and model-predicted values. A student's *t*-test for slope (H_o: slope=1) was evaluated at $P < 0.01$.

3. DATA

To conduct this study 3 datasets were used. Dataset 1 (real-world measurements) was comprised of *Bos Tarus* steers (n = 80) reported in the evaluation study of McPhee et al. (2014). Datasets 2 and 3 (n = 80) were artificially created from dataset 1 where dataset 2 had 3 mm added to each observed value and dataset 3 was rotated by multiplying the observed value by -1 and adding twice the mean of the observed value. Summary statistics for the data used in this study is reported in Table 2. The predicted BeefSpecs values (n = 80) using Dataset 1 as inputs were 4.7, 13.4, 9.61 and 1.86 for minimum, maximum, mean, and SD, respectively. The predicted values were used in each of the 3 evaluations.

Table 2. Summary of datasets used in the statistical evaluation of three P8 fat (mm) datasets.

	1	2	3
n	80	80	80
Minimum	6.00	9.00	3.35
Maximum	16.00	19.00	13.35
Mean	9.68	12.68	9.68
SD	2.21	2.21	2.21

4. RESULTS

The mean bias was under-predicted for all datasets (Table 3). There were significant differences ($P < 0.01$) in the mean bias of dataset 2 and no significant differences ($P > 0.05$) for datasets 1 and 3. There was no significant difference ($P > 0.05$) when testing for slope (H_o: slope = 1) for datasets 1 and 2 but significant differences ($P < 0.01$) for slope in dataset 3 (Table 3). Dataset 1 had the lowest RMSEP. The decomposition of the MSEP demonstrated that the majority of the error contained in the predictions was of a random nature for dataset 1. In dataset 2, the majority was in the associated bias and in dataset 3 it was in the slope. The MEF of 0.39 in dataset 1 was reasonable between the observed and predicted but the values < 0 for datasets 2 and 3 indicate a very poor agreement between observed and predicted values. A plot of the observed versus predicted final P8 fat with a 1:1 (y = x) line illustrates the relationship that each dataset has to the 1:1 line (Figure 1). Figure 2 illustrates the residuals (observed – predicted) with a horizontal line (y = 0) and the upper and lower control limit boundaries of 1.5mm; 65, 14, and 31% of the residuals line within ± 1.5mm for datasets 1, 2, and 3, respectively.

5. DISCUSSION AND CONCLUSIONS

Objections to the use of regression have been extensively outlined by Mitchell (1997). In brief four of the five objections are as follows: (1) the fraction of variation in the Y values explained by the regression (r^2) is of no relevance since it is not intended to make predictions from the fitted line.; (2) ambiguity in a null hypothesis test because the more scatter in the points, the greater the standard error of the slope and the smaller the computed value of the test statistic so that it is harder to reject the null hypothesis hence a paradoxical result that regressions from highly scattered samples of points are more likely to have slopes not significantly different from 1 or mean deviation significantly different from 0.; (3) the fitted line is irrelevant to validation because model validation is related to deviations from observed and model predicted values not the fitted line; and (4) violation of normal distribution e.g., the observations are values from either a series in time or space or are accumulated values or are auto-correlated and x values have error.

Table 3. Statistical evaluation across 3 datasets of differences between observed and BeefSpecs predicted final P8 fat.

Item	Datasets		
	1	2	3
n	80	80	80
Mean observed, mm	9.68	12.68	9.68
Mean predicted, mm	9.61	9.61	9.61
Mean bias, mm	0.06	3.06	0.06
MSEP[A]	2.96	12.33	13.57
Root-MSEP, mm	1.72	3.51	3.68
Bias, %	0.13	76.05	0.03
Slope, %	5.77	1.38	79.47
Random, %	94.10	22.57	20.51
Modelling Eff	0.39	-1.55	-1.80
Additional Statistics			
P^B	0.75	< 0.01	0.98
r^2	0.43	0.43	0.43
β_1 coefficient	0.78	0.78	-0.78
P^C	0.03	0.03	< 0.01

[A]MSEP = mean square prediction error, Bias = MSEP decomposed into error due to overall bias of prediction; Slope = MSEP decomposed into error due to deviation of the regression slope from unity, Random = MSEP decomposed into error due to the random variation
[B]Probability of paired t-test for the mean bias ($P < 0.05$)
[C]Probability of student's two-tailed t-test for the slope (H_o: slope=1) at ($P < 0.01$)

The message about misusing regression has been repeated on numerous occasions (Loague and Green, 1991; Mayer and Butler, 1993; Mitchell, 1997; Mitchell and Sheehy, 1997; Tedeschi, 2006) and the use of the simultaneous F-test applicable to deterministic models where model output has no error has been reported (Mayer et al., 1994). However, scientists and modellers continue to insist that r^2 and statistical tests of mean bias and slope need to be included when validating models. The fundamental issue is that scientists and modellers use the regression of best fit not the deviation of Y - X (observed – predicted) for model evaluation. The other issue that the scientific community struggles with is that MEF is a number between 1 and – infinity while r^2 is between 1 and 0. The reporting of r^2 is more highly regarded than the value it actually provides for discerning model quality. The illustration of the deviation in Figure 2 clearly drives the point home that r^2 does not mean that the model is a good fit. The decomposition of the MSEP into bias, slope and random components along with the reporting of deviations with an upper and lower control limit is highly recommended. Even though (3) to (5) use components of a regression to calculate the values the decomposition components add up to MSEP i.e., they are directly related to the MSEP that is universally accepted as the best method of reporting differences (Tedeschi 2006) between an observed and model predicted value.

In regards to the statistical test on the mean bias several authors have reported methods that they consider acceptable. For example Reckhow et al. (1990) suggests that a one-way t-test for the mean deviation being less than a specified value when the specification of the critical value is similar to the criterion of an envelope of acceptable precision could be applied. Tedeschi (2006) also makes the point that a paired t-test of the difference of the means is preferable to a t-test of the difference of the means since the former one is less conservative and removes any covariance between the data points.

In conclusion, the authors cannot emphasize strongly enough that model validation be conducted with the standard reporting of summary statistics, mean bias, RMSEP followed by the decomposition of MSEP (bias, slope, and random components) and the graphical display of deviations with upper and lower control limits.

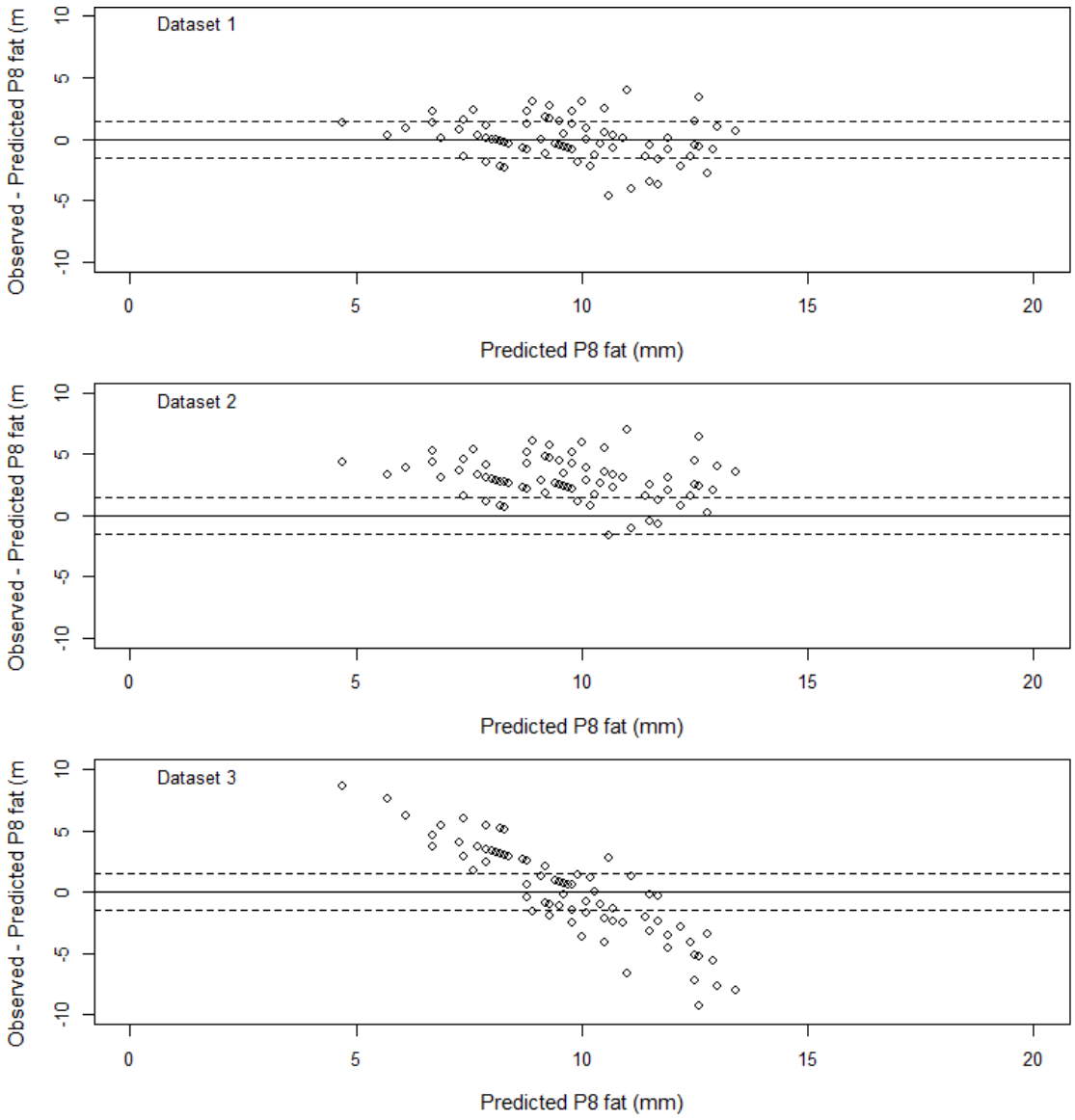

Figure 2. Differences of observed – predicted (residuals) versus predicted P8 fat (mm) with upper and lower control limits of 1.5 mm

ACKNOWLEDGMENTS

Funding from Meat and Livestock Australia who have supported both the development of BeefSpecs and the 3D cameras to assess hip height, P8 fat, and muscle score and hence the evaluation of observed and predicted or assessed values has continued to be evaluated. Data to conduct such studies could not have been possible without the assistance of a number of Local Land Service Officers (formerly NSW DPI Extension Officers) who have played key roles in the collection of industry data. In particular, Brett Littler and Jason Siddell are thanked.

REFERENCES

Bibby, J. and Toutenburg, H. (1977). *Prediction and improved estimation in linear models*. Akademie - Verlag Berlin, German Democratic Republic: John Wiley & Sons Ltd.

Graser, H.-U., Tier, B., Johnston, D.J. and Barwick, S.A. (2005) Genetic evaluation for the beef industry in Australia. *Australian Journal of Experimental Agriculture, 45*(8), 913-921. doi: https://dx.doi.org/10.1071/EA05075

Loague, K. and Green, R. E. (1991). Statistical and graphical methods for evaluating solute transport models: Overview and application. *Journal of Contaminant Hydrology, 7*(1-2), 51-73.

Mayer, D. G. and Butler, D. G. (1993). Statistical validation. *Ecological Modelling, 68*(1-2), 21-32.

Mayer, D. G., Stuart, M. A. and Swain, A. J. (1994). Regression of real-world data on model output: An appropriate overall test of validity. *Agricultural Systems, 45*(1), 93-104. doi:http://dx.doi.org/10.1016/S0308-521X(94)90282-8

McPhee, M. J., Walmsley, B. J., Mayer, D. G. and Oddy, V. H. (2014). BeefSpecs fat calculator to assist decision making to increase compliance rates with beef carcass specifications: evaluation of inputs and outputs. *Animal Production Science, 54*(12), 2011-2017. doi:http://dx.doi.org/10.1071/AN14614

Mitchell, P. L. (1997). Misuse of regression for empirical validation of models. *Agricultural Systems, 54*(3), 313-326.

Mitchell, P. L. and Sheehy, J. E. (1997). Comparison of predictions and observations to assess model performance: a method of empirical validation. In M. J. Kropff, P. S. Teng, P. K. aggarwal, J. Bouma, B. A. M. Bouman, J. J.W. and H. H. Van Laar (Eds.), *Applications of systems approaches at the field level* (pp. 437-451). Great Britain: Kluwer Academic Publishers.

Montgomery, D. C. (1991). *Introduction to Statistical Quality Control*. Singapore: John Wiley & Sons, Inc.

Pineiro, G., Perelman, S, Guerschman, J. P. and Paruelo, J. M. (2008). How to evaluate models: Observed vs. predicted or predicted vs. observed?. *Ecological Modelling*, 216, 316-322.

R Development Core Team, (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from URL http://www.R-project.org Downloaded on 4th August 2009.

Reckhow, K. H., Clements, J. T. and Dodd, R. C. (1990). Statistical evaluation of mechanistic water-quality models. *Journal of Environmental Engineering, 116*(2), 250-268.

Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. *Agricultural Systems, 89*(2-3), 225-247.

Theil, H. (1966). *Applied economic forecasting*: North-Holland Pub. Co.

Upton, W., Donoghue, K., Graser, H. and Johnston, D. (1999). *Ultrasound proficiency testing.* Paper presented at the Association for the Advancement of Animal Breeding and Genetics.