Model evaluation: The misuse of statistical techniques when evaluating observations versus predictions

Malcolm McPhee[1*], Jonathan Richetti[2], Barry Croke[3], and Brad Walmsley[1,4]

[1] NSW Department of Primary Industries and Regional Development, Livestock Industries Centre, University of New England, Armidale, NSW, Australia
[2] CSIRO, Floreat, WA, Australia
[3] Mathematical Sciences Institute and Fenner School of Environment & Society, The Australian National University, ACT, Australia
[4] Animal Genetics and Breeding Unit (AGBU), a joint venture of NSW Department of Primary Industries and Regional Development and University of New England, Armidale, NSW, Australia

Abstract

Mathematical modellers, decision support developers, statisticians, and students evaluate the differences between observed and model predicted values. When evaluating models, it is far too easy to conduct model evaluation by fitting a linear regression to the data. In this paper, steps are presented on 'how to' evaluate a model using deviance metrics rather than reporting r^2 from fitting a linear regression. The paper aims to provide sound reasoning, with data, against using r^2. The paper addresses five arguments, previously put forward, for not fitting a linear regression when conducting model evaluation: i) Misapplication of regression; ii) Ambiguity of null hypothesis tests; iii) Lack of sensitivity; iv) Fitted line is irrelevant to validation; and v) Violation of regression assumptions. Statistical, deviance, and quality control metrics are outlined. Three models using the BeefSpecs drafting tool are reported in this paper. Each model (n = 80) had an r^2 of 0.43. A mean bias of 0.06, -2.90, and -0.11 mm, and a root mean square error of prediction (RMSEP) of 1.72, 3.37, and 3.70 mm for models 1, 2, and 3, respectively. A modelling efficiency (MEF) of 0.39, -1.34, and -1.83, and 91, 51, and 56% of predictions within upper and lower quality control limits for models 1, 2, and 3, respectively. These metrics highlight the pitfall of reporting r^2 from using regression. Minimum recommended steps of 'how to' conduct model evaluation are: a plot of the residuals with quality control limits and a table of metrics including mean observed, predicted and bias, RMSEP, and MEF.

Keywords

Deviance metrics; modelling efficiency; bias; slope; deviance

Code & Data availability

The BeefSpecs drafting tool used to generate the predictions reported in this paper can be found at https://beefspecs.agriculture.nsw.gov.au/drafting. The recommended metrics here presented are available via: https://github.com/JRichetti/model_evaluation.

Correspondence:
Contact M. McPhee at malcolm.mcphee@dpi.nsw.gov.au

Cite this article as:
McPhee, M., Richetti, J., Croke, B., and Walmsley, B.
Model evaluation: The misuse of statistical techniques when evaluating observations versus predictions
Socio-Environmental Systems Modelling, vol. 6, 18758, 2024, doi:10.18174/sesmo.18758

Socio-Environmental Systems Modelling
An Open-Access Scholarly Journal
http://www.sesmo.org

1. Introduction

Evaluation of observations *versus* predictions from models is a common practice used to assess the accuracy and precision of models. Mathematical modellers, decision support developers, statisticians, and their students from environmental, hydrology, spatial, cropping, and livestock disciplines often use a range of techniques to evaluate models. When evaluating models, the presentation of results is frequently reported where several statistical methodologies are used. For example, the coefficient of determination (r^2) from a regression fitted to the observed *versus* predicted data is commonly reported in journal publications, conference proceedings, and when presenting results at conferences (R^2 is also a commonly used symbol instead of r^2, and there are also multiple definitions of R^2 in use which can confuse readers (Kvålseth, 1985)). In general, deviance metrics are also reported alongside an r^2 and plots of observed *versus* predicted are frequently reported with a 1:1 line (i.e., y=x line). The misuse of regression for empirical validation of models has previously been reported by Kvålseth (1985) and Mitchell (1997). However, regression continues to be the most dominant statistical methodology used in model evaluation.

Techniques used in model evaluation are the focus of this paper with an overall aim of developing 'how to' steps for good model evaluation. The concept of "good modelling practice" has already been considered (Jakeman et al. 2006). Jakeman et al. (2006) illustrated ten iterative steps for good modelling practice:

- Definition of the purposes for modelling.
- Specification of the modelling context scope and resources.
- Conceptualisation of the system, specifications of data and other prior knowledge.
- Selection of model features and families.
- Choice of how model structure and parameter values are to be found.
- Choice of estimation performance criteria and technique.
- Identification of model structure and parameters.
- Conditional verification including diagnostic checking.
- Quantification of uncertainty.
- Model evaluation or testing (other models, algorithms, comparisons with alternatives).

It could be argued that model evaluation or testing, the 10th step, is the most critical step to be undertaken where the worthiness of all the other steps is determined. Model evaluation assesses the impact of the model and its usefulness, and where possible independent observations are used to make the evaluation. It is the 10th step of good modelling practice that this paper addresses. Therefore, the aim of this paper is to outline several steps that constitute good model evaluation practice. As Jakeman et al. (2006) stated "sustained attention needs to be made to ongoing improvements in developing techniques to provide credibility and integrity to the models developed". Ongoing improvements to methodologies applied to model evaluation also need to be undertaken. The ongoing process of striving to improve model evaluation raises questions, such as: Have we introduced too much complexity into model evaluation? Is there an alternative approach to how models get evaluated? And if so, what are the steps required to achieve a rigorous evaluation?

Several techniques and methodologies for evaluating models have previously been reported. Tedeschi (2006) reviewed several techniques, Bellocchi et al. (2010) reviewed the issues and methodologies of validating biophysical models and Bennett et al. (2013) characterised the performance of environmental models that includes a list of quantitative methods that could be used in evaluating models. All three reviews by Tedeschi (2006), Bellocchi et al. (2010), and Bennett et al. (2013) encourage modellers to use regression in quantifying models. The list of techniques outlined by Bennett et al. (2013) illustrates the length and level of complexity that modellers have been striving to achieve when quantifying models. In some cases, calculations to quantify observed *versus* predicted values are similar but use different names e.g., the Nash-Sutcliffe model efficiency (NSE also referred to as R^2) (Nash & Sutcliffe, 1970) is the same calculation as the modelling efficiency (MEF) deviance metric developed by Loague & Green (1991) and reported by Mayer & Butler (1993). Thus, an overwhelming number of metrics, including some with different names, are available but an understanding of how to adequately use them is lacking. The MEF notation is used in this paper.

The paper by Mitchell (1997) provided sound reasoning against using regression for validating models and suggested alternatives (Mitchell, 1997; Mitchell & Sheehy, 1997). Mitchell (1997) outlined five objections to using regression:

i) Misapplication of regression. The fraction of variation in the Y values (observed) explained by the regression (r^2) is of no relevance since it is not intended to make predictions from the fitted line.

ii) Ambiguity of null hypothesis tests. Ambiguity exists 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. Therefore, 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.

iii) Lack of sensitivity. Regression lacks sensitivity in model evaluation because distinguishing the points from a random cloud is rarely necessary at the final stages of model development.

iv) Fitted line is irrelevant to validation. The fitted line is irrelevant to validation because model validation is related to deviations from observed and model predicted values not the fitted line.

v) Violation of regression assumptions. Violation of regression assumptions (i.e., homogeneous variance along the x-axis and with the x and y data values as well as the residuals being normally distributed) e.g., the observations are values from either a series in time or space, or are accumulated values, or are autocorrelated and X values (predictions) have error.

Several authors (Loague & Green, 1991; Flavelle, 1992; Reckhow et al., 1992) as reported by Mitchell (1997) have stated that there are benefits from a regression because it provides an "objective and quantitative method for evaluating models". Regression is also a familiar technique and is considered an easy option. Therefore, the easy option to evaluate models objectively and quantitatively is a regression of the observed *versus* predicted. Even though Loague and Green (1991), Flavelle (1992), and Reckhow et al. (1992) all encourage the use of regression they do however, all acknowledge that problems do exist in satisfying the regression assumptions. Tedeschi (2006), Bellocchi et al. (2010), and Bennett et al. (2013) also lean towards including the results of a regression in model evaluation, but they do recommend checking the data and conducting visual assessments. For instance, Bennett et al. (2013) indicated there may be some cases where a simple graphical representation of output is sufficient.

Many research scientists and students have been led astray by: (a) being familiar with regressions but failing to understand the underlining assumptions; and (b) the demands of journal and conference proceeding editors suggesting that r^2 should be included in the results of model evaluation. Therefore, developing a rigorous alternative set of 'how to' steps to produce a good modelling evaluation practice has merit.

This paper aims to provide guidelines for quantitative model evaluation without relying on the r^2 from fitting linear regressions to observed *versus* predicted data. Therefore, the objectives of this paper are to: (1) review the five objections of Mitchell (1997) and provide additional quantitative detail to their reasoning as required; (2) provide illustrations using observed *versus* predicted data from a published model to illustrate that model evaluation can be achieved without using regression; and (3) provide 'how to' steps to quantify model evaluation without using regression and reporting r^2.

The word validation is often used when evaluating models, but validation can mean different things to those in different disciplines. The meaning of verification, validation, and confirmation when evaluating models has been discussed by Oreskes et al. (1994). In this paper 'model evaluation' is used throughout rather than verification, validation, or confirmation.

The published model used in this paper is the BeefSpecs drafting tool (Walmsley et al., 2011); the mathematical models and equations have been reported by Walmsley et al. (2014). In brief, the BeefSpecs drafting tool has been developed for on-farm drafting to allow beef producers to explore management changes to meet market specifications. Producers are penalized if they do not meet stringent market specifications related to fat distribution (i.e., subcutaneous fat P8 rump fat thickness (P8 fat, mm) or 12th-rib fat thickness (12th-rib fat, mm) sites) and hot standard carcass weight (kg). Improving market compliance rates by assisting producers to meet market specifications was estimated to be worth well over AU$51 million/year to the Australian beef industry (Lollback, 2012) and even more when a reduction in feeding costs is taken into consideration. Preliminary results, using the BeefSpecs on-farm drafting tool, on the misuse of regression when comparing observed *versus* predictions of final P8 fat (mm) were published by McPhee & Walmsley (2017) at the 2017 International Congress on Modelling and Simulation (MODSIM2017).

2. Model evaluation techniques

There are several model evaluation techniques that are used. Mayer & Butler (1993) describe four main categories: subjective assessment, visual techniques, deviance measures (i.e., deviance metrics) and statistical tests (i.e., statistical metrics). Several mathematical notations are used throughout this paper and are defined in Table 1. This description of model evaluation techniques has been categorised into three metrics: statistical metrics, deviance metrics, and quality control metrics.

Table 1: Mathematical notation and description.

Notations	Description	Metric		
		Statistical	Deviance	Quality control
Deviation	The difference between observed and model predicted values $(Y_i - f(X_1,...,X_p)_i)$		✓	
$f(X_1,...,X_p)_i$	The i[th] model predicted (or simulated) value	✓	✓	
$\bar{f}(X_1,..., X_p)_i$	Mean of model predicted (or simulated) values	✓	✓	
LCL	Lower control limit			✓
MSEP	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	✓		
Deviance	MSEP decomposed into error due to the deviance variation	✓		
MAE	Mean absolute error		✓	
MA%E	Mean absolute percent error		✓	
MEF	Modelling efficiency	✓		
NSE	Nash-Sutcliffe model efficiency	✓		
r^2	Coefficient of determination		✓	
RMSEP	Root mean square error of prediction	✓		
SSR	Sums of squares of regression about the fitted line	✓		
SSE	Sums of squares of the error	✓		
SSTO	Sums of squares total	✓		
UCL	Upper control limit			✓
WCL	Within lower and upper control limits			✓
$X_1,...,X_p$	The p[th] input to the model	✓	✓	
Y_i	The i[th] observed or measured value	✓	✓	
$\bar{Y}$	Mean of the observed (or measured) values	✓		
$\hat{Y}_i$	The i[th] linear value of the regression fitted through the data	✓		

2.1 Statistical metrics

The following statistical metrics have been used in this paper. Linear regression (1) is commonly used to evaluate a model:

$$Y_i = \beta_o + \beta_1 \times f(X_1, \ldots, X_p)_i + \varepsilon_i, \quad (1)$$

where β_0 and β_1 are the regression parameters for the intercept and slope, respectively and ε_i is the i[th] deviance error assumed to be from a single population that is independent and normally distributed ~ $\mathrm{N}(0, \sigma^2)$.

The decomposition of mean square error of prediction (MSEP) (2), also referred to as the mean square error (MSE), was first introduced by Theil (1966) and outlined with additional explanation by Bibby & 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) + Deviance (5). Both the slope and deviance components represent the sample variance of predicted and observed values.

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

The bias, slope, and deviance components are generally reported as percentages of the total MSEP:

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

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

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

Modelling Efficiency (MEF) (6), like the NSE (Nash & Sutcliffe, 1970), is described by Loague & Green (1991), and reported by Mayer & Butler (1993) as a dimensionless statistic that directly relates model predictions to observed data:

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

The r^2 for a linear regression (7) is interpreted as the proportion of variation explained by the fitted line:

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

where

$$\sum_{i=1}^{n} \hat{Y}_i \times \varepsilon_i = 0 \quad (8)$$

and

$$\sum_{i=1}^{n} \varepsilon_i = 0 \quad (9)$$

2.2 Deviance metrics

Mayer & Butler (1993) explain what is meant by deviance metrics used for validation and outline some of the pitfalls. In brief, deviance metrics are calculations of the difference between observed and predicted values paired to time, location, treatment, etc. The mean absolute error (MAE) (10) and mean absolute percent error (MA%E) (11) (Shaeffer, 1980) are often reported. The MSEP (2) is a common deviance metric that compares the observed values *versus* the predicted values; also referred to as second moments (Picard & Cook, 1984). It is worth noting that there is a linear relationship between MSEP (2) and MEF (6), so these should not be considered independent measures of model performance. The square root of the MSEP (RMSEP) (12) is generally reported rather than the MSEP (2), also referred to as the root mean square error (RMSE). The MAE and RMSEP are in the same units as the data and are therefore a meaningful metric to report. Hodson (2022) stated that neither the MAE or RMSEP is the better metric, however, Hodson (2022) reports that the MAE is optimal for Laplacian errors and RMSEP more suited for normal (Gaussian) errors.

$$MAE = \frac{\sum_{i=1}^{n}(|Y_i - f(X_1,\ldots,X_p)_i|)}{n} \quad (10)$$

$$MA\%E = 100 \times [\sum_{i=1}^{n}(|Y_i - f(X_1,\ldots,X_p)_i|/|Y_i|)]/n \quad (11)$$

$$RMSEP = \sqrt{MSEP} \quad (12)$$

2.3 Quality control metrics

Quality control metrics are used extensively in engineering (Montgomery, 1991). Upper and lower control limits displayed on Shewhart quality control charts (Shewhart Control Charts, 2000) provide tolerance levels for technicians to flag that a manufactured product is out of tolerance. The Shewhart control chart measures a quality characteristic, say *w*, and the mean of *w* is μ_w (Montgomery, 1991). The general model of a Shewhart control chart is as follows:

$$UCL = \mu_w + k\sigma_w, \quad (13)$$

$$Centre\ line = \mu_w, \quad (14)$$

$$LCL = \mu_w - k\sigma_w, \quad (15)$$

where UCL is the upper control limit, LCL is the lower control limit, and *k* is usually chosen to be 3, where 99.7% of the observed data of a normal distribution lies within 3σ, 95% lies within 2σ, and 68% lies within 1σ (Sokal & Rohlf, 1995). The smaller the *k* the tighter the control limits, but with lower confidence.

When dealing with observed *versus* predicted values the centre line (14) = 0. Each discipline should be able to provide acceptable limits for the model differences between observed and predicted. Mitchell & Sheehy (1997) recommend that the UCL and LCL are determined before model evaluation is undertaken. The BeefSpecs drafting tool and the underlying models to predict final P8 fat (mm) along with the observed ultrasound P8 fat (mm) provide the data to illustrate the misuse of regression. For example, to be accredited as an ultrasound technician to scan cattle requires a level of proficiency for assessing fat depth to be within 1.5 mm of the mean (i.e., an assessor considered to have high level skills in assessing fat) (Upton et al., 1999). However, the UCL and LCL errors for model evaluation are larger when errors associated with the model are taken into consideration. The UCL and LCL when comparing observed *versus* prediction of P8 fat are based on a sensitivity analysis conducted in a BeefSpecs evaluation of inputs and outputs (McPhee et al., 2014). The sensitivity analysis study of BeefSpecs found that "the average sensitivity of animals across sexes and frame scores with an initial LW of 200 kg and initial P8 fat of 2 mm was 1.51 mm/mm. This result means that an error in the estimation of initial P8 fat of 2mm will result in an error of up to 3 mm in the prediction of final P8 fat". Therefore, the UCL and LCL were set at 2 x 1.51 mm = 3.02 mm and rounded to 3 mm.

3. Data

To conduct this study three cattle models were simulated and evaluated against 80 observed values of final ultrasound P8 fat (mm). Models 1 to 3 were comprised of *Bos Tarus* steers. Summary statistics of final ultrasound P8 fat (mm) of observed and BeefSpecs drafting tool predictions for models 1 to 3 are reported in Table 2.

Table 2: Summary statistics of observed ultrasound assessments and BeefSpecs drafting tool predictions for models 1 to 3 of final P8 fat (mm) of cattle used in the statistical evaluation of three models.

	Observed	Predicted		
		Models		
		1	2	3
n	80	80	80	80
Minimum	6.00	4.70	7.70	5.96
Maximum	16.00	13.40	16.40	14.66
Mean	9.68	9.61	12.58	9.78
SD	2.21	1.86	1.85	1.87

4. Results

The density distribution of the observed and predicted values and the differences between the observed and predicted values of final P8 fat (mm) are reported in Figures 1 and 2, respectively.

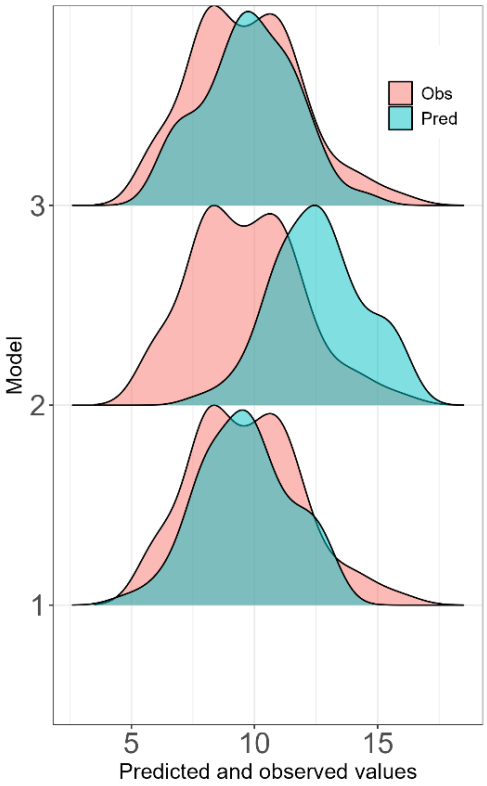

Figure 1: Density distribution of observed (Obs) and final P8 fat (mm) predictions (Pred) of models 1 to 3.

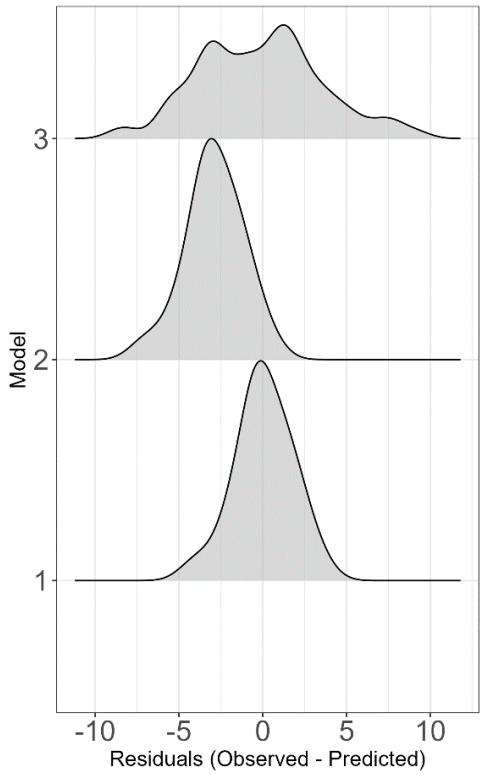

Figure 2: Density distribution of difference (observed - final P8 fat (mm) predictions) of models 1 to 3.

The mean bias shows that all models under-predicted the observations even though only slightly in some cases (Table 3). There were significant differences ($P < 0.01$) in the mean bias ($\mu_1 \neq \mu_2$) of model 2 and no significant differences ($P > 0.05$) for models 1 and 3. There were no significant differences ($P > 0.05$) when testing for slope (H_o: slope = 1) for models 1 and 2 but significant differences ($P < 0.01$) for slope in model 3. Model 1 had the lowest RMSEP of 1.72 mm. The decomposition of the MSEP demonstrated that most of the error contained in the predictions was due to disturbances (i.e., deviance) for model 1. In model 2, the majority was in the associated bias and in model 3 it was in the slope. The MEF of 0.39 in model 1 indicated reasonable agreement between the observed and predicted final P8 fat but the MEF < 0 for models 2 and 3 indicated poor agreement between observed and predicted values. Mayer & Butler (1993) stated that "any model giving a negative value cannot be recommended". The quality control metric of UCL = 3 mm and LCL = -3 mm revealed that 91% of the residuals were within the upper and lower controls of model 1 and the residuals of models 2 and 3 were both <=56%. Model 1 had the lowest MAE of 1.34 mm and lowest MA%E of 15%. The r^2 was the same across all models and the β_1 coefficient was the same for all models but the β_1 coefficient was negative for model 3 rather than positive. Highly significant differences ($P < 0.01$) for slope (H_o: slope=1) were detected for model 3 but models 1 and 2 revealed a tendency ($P < 0.05$) (Table 3).

Table 3: Statistical evaluation of final P8 fat (mm) across 3 models of observed and predicted values using the BeefSpecs drafting tool.

Item	Models		
	1	2	3
n	80	80	80
Mean observed, mm	9.68	9.68	9.68
Mean predicted, mm	9.61	12.58	9.78
Mean bias, mm	0.06	-2.90	-0.11
MSEP[A]	2.96	11.35	13.70
Root-MSEP, mm	1.72	3.37	3.70
Bias, %	0.13	74.21	0.09
Slope, %	5.77	1.38	79.83
Deviance, %	94.10	24.41	20.09
Modelling Eff	0.39	-1.34	-1.83
WCL[B], %	91	51	56
MAE[C], mm	1.34	2.94	3.00
MA%E[D], %	15	34	33
Additional Statistics			
P^E	0.75	< 0.01	0.88
r^2	0.43	0.43	0.43
β_1 coefficient	0.78	0.78	-0.78
P^F	0.03	0.03	< 0.01

[A]MSEP = mean square error of 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, Deviance = MSEP decomposed into error due to the deviance variation.
[B]WCL = within upper and lower control limits.
[C]MAE = mean absolute error (Shaeffer, 1980).
[D]MA%E = mean absolute percent error (Shaeffer, 1980).
[E]Probability of paired t-test for the mean bias ($P < 0.05$).
[F]Probability of student's two-tailed t-test for the slope (H_o: slope=1) at ($P < 0.01$).

A plot of the observed *versus* predicted final P8 fat with a 1:1 ($y = x$) line illustrates the relationship that each model has to the 1:1 line (Figure 3).

Figure 4 illustrates the residuals (observed – predicted) with a horizontal line (y = 0) and the upper and lower control limit boundaries of 3.0 mm. The within control limits (± 3.0 mm) percentages are 91, 51, and 56% for models 1, 2, and 3, respectively (Table 3).

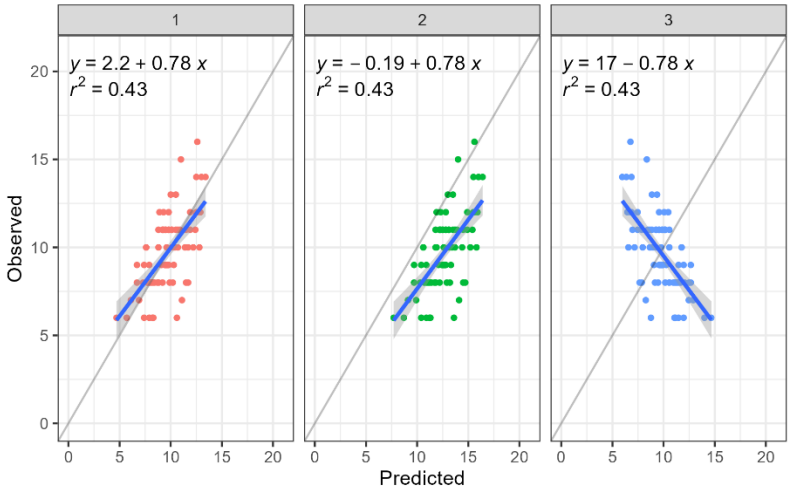

Figure 3: Scatter plots (enclosed circles) of models 1 to 3 of final P8 fat (mm) (observed *versus* predicted) where the solid black line (thin) is the 1:1 relationship and the blue solid line (thick) is the regression line. The regression equation and r^2 are reported for each model to demonstrate the misuse of regression for empirical validation of models.

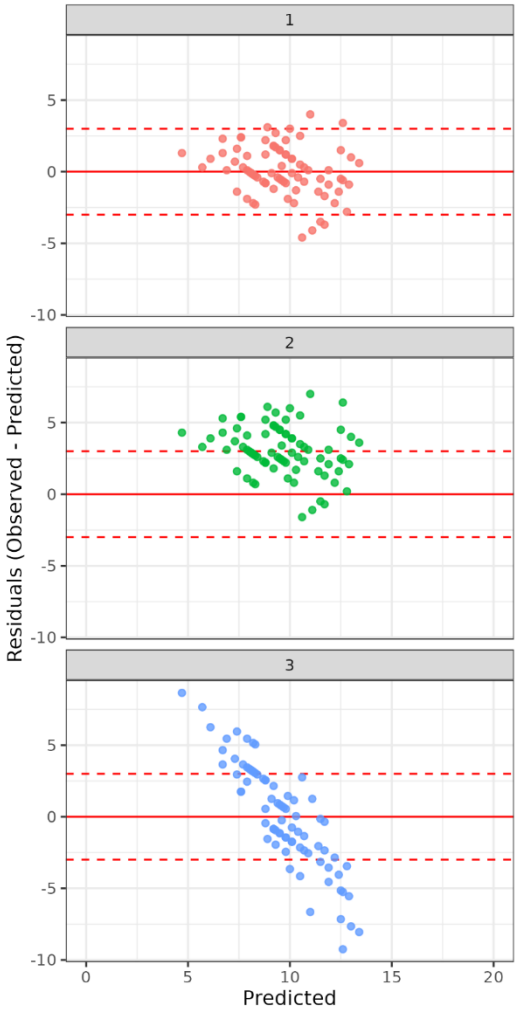

Figure 4: Residuals (enclosed circles; observed – predicted) *versus* predicted P8 fat (mm) with upper and lower control limits (dashed lines) of 3.0 mm and solid line residuals=0.

5. Discussion

Concerns about misusing regression when conducting model evaluation have been repeated on numerous occasions (Kvålseth, 1985; Loague & Green, 1991; Mitchell, 1997; Mitchell & Sheehy, 1997; Tedeschi 2006) and the use of the simultaneous F-test applicable to deterministic models has been reported by Mayer et al. (1994). The fundamental issue is that many scientists and modellers use the regression of best fit not the deviance of the observed – predicted (residuals) for model evaluation. Table 3 reports the statistical metrics where the r^2 and the slope coefficient are the same for models 1 and 2 and the β_1 coefficient for model 3 reverses the inequality of models 1 and 2 (-1 x β_1). Figure 4 illustrates the deviation across all three models and that the r^2 reported in Figure 3 and Table 3 does not mean that the model is a good fit. This highlights the pitfall of relying on r^2 to evaluate models. The decomposition of the MSEP (2) into bias, slope, and deviance components along with the reporting of deviations with an upper and lower control limit is highly recommended. Even though (4) and (5) use components of a regression to calculate the values, the decomposition components (deviances) (3) to (5) add up to the 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. Regarding the statistical test on the mean bias, several authors have reported methods that they consider acceptable. For example, Reckhow et al. (1992) 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 like the criteria of an envelope of acceptable precision. Tedeschi (2006) also states "that a paired t-test is preferable to a t-test of the difference of the means since the former paired t-test is less conservative and removes any covariance between the data points".

This paper provides a quantitative response to the five objections to using regression for validating models that Mitchell (1997) outlined:

i) Misapplication of regression

Developing regressions to conduct model evaluations, using ordinary least squares to make predictions of *y* from *x*, based on assumptions (Draper & Smith, 1966; Sokal & Rohlf, 1995), is commonly used by research scientists and students from a range of disciplines. In fact, one could argue that it is so familiar, as stated by Flavelle (1992), that it is the most common use of regression by research scientists and students. The key point that Mitchell (1997) is making here is that the fraction of variation in the *y* values (observed) explained by the regression (r^2) is of no relevance since it is not intended to make predictions from the fitted line. Mitchell (1997) highlights that the least squares method employed by regression sets out to minimise the variation between the *y*- and *x*-axis and when comparing observed *versus* predictions minimising the variation is not the object of the comparison but rather the variation in the relationship to the 1:1 line is what is being assessed. This is why the MEF metric (5) is different from an r^2 in that it quantitatively emphasizes the 1:1 line relationship. However, many in the scientific community struggle with stating the MEF because it gives a number between $-\infty$ and 1. It may seem harsh with values < 0 (Table 3; models 2 and 3) but the MEF does quantify a model that takes into consideration the 1:1 line. The RMSEP and MAE are the highest across models 2 and 3 (Table 3) and quantitatively back up the MEF as a metric that can provide guidance on whether the model predictions are accurate in relation to the observed values.

ii) Ambiguity of null hypothesis tests

The t- or F-tests are frequently conducted in model evaluations on the mean bias ($\mu_1 \neq \mu_2$) and the slope (H_o: slope=1). It could also be conducted on the intercept; "testing that the intercept does not differ from zero" (Mitchell, 1997) but this is rarely reported. Sample size and the associated scatter with a large sample size is an issue here; "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. This leads to the paradoxical result that regressions from highly scattered samples of points are more likely to have slopes not significantly different from 1!" (Mitchell, 1997). "The test can fail either because the slope is really not different from 1 or because there is much scatter around the line" (Mitchell, 1997). The models in this paper all have the same sample size (n = 80) therefore, the point raised by Mitchell (1997) cannot be illustrated. However, the $P = 0.03$ for models 1 and 2, both show a tendency ($P < 0.05$) rather than a highly significant ($P < 0.01$) difference, thus inferring that the t-test is a stringent test on the slope when sample sizes are larger. Based on the predefined UCL and LCL of ± 3 mm, 91% of the data in model 1 were within the control limits as opposed to models 2 and 3, which were <= 56%. For the prediction of final P8 fat (mm), the results of 91% is considered adequate. However, the level of adequacy of a model will depend on user engagement to determine acceptable levels.

iii) Lack of sensitivity

Mitchell (1997) reports that once model evaluation is reached, when using regression, there is in general good agreement with observations and predictions. In other words, it is a trivial step and thus predictable that a regression will fit the data. Sample size and the spread of data play a role in the lack of sensitivity in the data. Mitchell (1997) attached an appendix to their publication indicating that a regression is not good enough to quantify how good the line of best fit is, once it has gone past the conventional thresholds of $P = 0.05$, 0.01, or 0.001. One could equally argue that low sample sizes e.g., $n < 7$, will also lack sensitivity. Low sample sizes impact all metrics outlined in this paper and thus conducting model evaluation when the sample size is small should be avoided. Generally, sample sizes >= 15 are considered acceptable. However, collecting observed data to evaluate complex systems can be expensive and difficult to achieve. Huth & Holzworth (2005) have developed a system called 'sensibility tests' that evaluate model usefulness. In 'sensibility tests' the model evaluation is made against more subjective, local experts feeling for model behaviour (Huth & Holzworth, 2005). The 'sensibility tests' of Huth & Holzworth (2005) are like the user predefining the lower and upper control limits on the plots of the residuals (Figure 4). The plots of the residuals are often easier to understand for non-modellers.

iv) Fitted line is irrelevant to validation

The fitted line is the best summary of a straight-line relationship (Mitchell, 1997) among the sample points of observed *versus* predictions in model evaluation. Therefore, the fitted line is irrelevant to validation. Models 1 to 3, illustrated in Figure 3, all having the same r^2 emphasizes that the fitted line is irrelevant. The deviations between observed and predicted as shown in Figure 4 with UCL and LCL on the residuals is highly recommended as a way forward. As stated above, the residuals are a lot easier for non-modellers to understand.

v) Violation of regression assumptions

The assumptions (Draper & Smith, 1966; Sokal & Rohlf, 1995) of linear regression are as follows:

- x (i.e., predicted) values are known without error.
- y (i.e., observed) values
    - should be a random sample.
    - independent of one another with common homogeneous variance along the x-axis and with residuals normally distributed. For example, if the observations are values from either a series in time or space, or are accumulated values, or are autocorrelated in any other way then the assumption of independence is suspect (Mitchell, 1997).

The first assumption is true for deterministic models provided the predictions are on the x-axis (Mayer et al., 1994). The density distributions of the residuals in Figure 2 demonstrates Gaussian, Gamma, and Uniform distributions for models 1 to 3, respectively.

The five objections of Mitchell (1997) and quantitative results, from using the BeefSpecs drafting tool, have delineated reasons why alternatives to the r^2 in model evaluations need to be implemented into good model evaluation practice. Model evaluation is the 10th step in good modelling practice (Jakeman et al., 2006).

The recommended 'how to' steps of good model evaluation are as follows:

1. Determine the UCL and LCL quality control limits to be displayed on a plot of the residuals (observed – predicted).
2. Plot observations *versus* predictions with a 1:1 line and a plot of the residuals (observed – predicted) with UCL and LCL quality control limits.
3. Table the mean observed, mean predicted, mean bias, MSEP, RMSEP, MEF, the decomposition of the MSEP (bias, slope, and deviance), and the percentage of residuals that lie within the UCL and LCL.
4. As a minimum:
    a. Plot residuals (observed – predicted) with UCL and LCL quality control limits.
    b. Table of metrics including:
        i. Mean observed,
        ii. Mean predicted,
        iii. Mean bias,
        iv. RMSEP, and
        v. MEF.

No model or model evaluation is perfect but, this paper highlights that alternative approaches to using the r^2 can be used when conducting model evaluation. The 'how to' steps of good model evaluation have been outlined in this paper. Furthermore, recommendations are made that are relevant and important for machine and deep learning algorithms. Modern algorithms can overfit and finding adequate 'how to' steps for model evaluation can be challenging. The 'how to' steps, as demonstrated in this paper are recommended to be included in model evaluation alongside cross-validation techniques for modern machine and deep learning algorithms (Richetti et al., 2023).

6. Conclusion

The misuse of regression still frequently occurs in model evaluation, and the authors of this paper acknowledge that we have also fallen into the misuse of regression either from failing to see the error of our ways (e.g., using the metrics inappropriately) or from falling into line with journal editors who require authors to use regression, to report results for publication. The authors agree with Kvålseth (1985); Tedeschi (2006); and Hodson (2022) and acknowledge that there is no one perfect quantitative metric for model evaluation but rather model evaluation should include several metrics. The authors minimum 'how to' steps recommendation is a report with a plot of the residuals with quality control limits and a table of metrics including mean observed, predicted and bias, RMSEP, and MEF.

Acknowledgements

Funding from Meat and Livestock Australia has supported both the development of the BeefSpecs drafting tool and the 3D camera technology to assess hip height, P8 fat, and muscle score and hence the evaluation of observed and predicted or assessed values has continued. Data to conduct such studies could not have been possible without the assistance of New South Wales Local Land Service Officers who have played critical roles in the collection of such industry data.

References

Bellocchi, G., Rivington, M., Donatelli, M., & Matthews, K. (2010). Validation of biophysical models: issues and methodologies. A review. Agronomy for Sustainable Development, 30(1), 109-130. https://doi.org/10.1051/agro/2009001

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., & Andreassian, V. (2013). Characterising performance of environmental models. Environmental Modelling & Software, 40, 1-20. https://doi.org/10.1016/j.envsoft.2012.09.011

Bibby, J., & Toutenburg, H. (1977). Prediction and improved estimation in linear models. John Wiley & Sons, Germany. (German)

Draper, N., & Smith, H. (1966). Applied regression analysis. John Wiley & Sons, New York.

Flavelle, P. (1992). A quantitative measure of model validation and its potential use for regulatory purposes. Advances in Water Resources, 15(1), 5-13. https://doi.org/10.1016/0309-1708(92)90028-Z

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development, 15(14), 5481-5487. https://doi.org/10.5194/gmd-15-5481-2022

Huth, N. I., & Holzworth, D. P. (2005). Common Sense In Model Testing In Zerger, A. and Argent, R.M. (eds) MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2005, Melbourne, Victoria, Australia. pp. 2804-2809. https://www.mssanz.org.au/modsim05/papers/huth.pdf

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software, 21(5), 602-614. https://doi.org/10.1016/j.envsoft.2006.01.004

Kvålseth, T. O. (1985). Cautionary Note about R^2. The American Statistician, 39(4), 279-285. https://doi.org/10.1080/00031305.1985.10479448

Loague, K., & Green, R. E. (1991). Statistical and graphical methods for evaluating solute transport models: Overview and application. Journal of Contaminant Hydrology, 7(1), 51-73. https://doi.org/10.1016/0169-7722(91)90038-3

Lollback, D. (2012). Livestock Data Link – linking supply chain partners. Meat & Livestock Australia. https://www.mla.com.au/globalassets/mla-corporate/research-and-development/program-

areas/ldl/documents/livestock-data-link---program-overview-cattle---linking-supply-chain-partners.pdf. Accessed January 27, 2023.

Mayer, D. G., & Butler, D. G. (1993). Statistical validation. Ecological Modelling, 68(1), 21-32. https://doi.org/10.1016/0304-3800(93)90105-2

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

McPhee, M. J., & Walmsley, B. J. (2017). Misuse of coefficient of determination for empirical validation of models. In Syme, G., Hatton MacDonald, D., Fulton, B. and Piantadosi, J. (eds) MODSIM2017, 22nd International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2017, Hobart, Tasmania. ISBN: 978-0-9872143-7-9. pp. 230–236. http://www.mssanz.org.au/modsim2017/B1/mcphee.pdf

McPhee, M. J., Walmsley, B. J., Mayer, D. G., & 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(11-12), 2011-2017. https://doi.org/10.1071/AN14614

Mitchell, P. L. (1997). Misuse of regression for empirical validation of models. Agricultural Systems, 54(3), 313-326. https://doi.org/10.1016/S0308-521X(96)00077-7

Mitchell, P. L., & 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 (Eds.), Applications of systems approaches at field levels (pp. 437-451). Kluwer Academic Publishers.

Montgomery, D. C. (1991). Introduction to Statistical Quality Control. John Wiley & Sons, New York.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology, 10(3), 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Oreskes, N., Shrader-Frechette, K., & Beltiz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. Science, 263, 641-646.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. Journal of the American Statistical Association, 79(387), 575-285. https://doi.org/10.1080/01621459.1984.10478083

Reckhow, K. H., Clements, J. T., & Dodd, R. C. (1992). Statistical Evaluation of Mechanistic Water-Quality Models. Journal of Environmental Engineering, 118(1), 155-156. https://doi.org/10.1061/(ASCE)0733-9372(1992)118:1(155.2)

Richetti, J., Diakogianis, F. I., Bender, A., Colaço, A. F., & Lawes, R. A. (2023). A methods guideline for deep learning for tabular data in agriculture with a case study to forecast cereal yield. Computers and Electronics in Agriculture, 205, 107642. https://doi.org/10.1016/j.compag.2023.107642

Shaeffer, D. L. (1980). A model evaluation methodology applicable to environmental assessment models. Ecological Modelling, 8, 275-295. https://doi.org/10.1016/0304-3800(80)90042-3

Shewhart Control Charts. (2000). In P. M. Swamidass (Ed.), Encyclopedia of Production and Manufacturing Management , pp. 685-686. Springer US. https://doi.org/10.1007/1-4020-0612-8_874

Sokal, R., & Rohlf, F. (1995). Biometry: the principles and practice of statistics in biological research (Vol. 3rd Edition). W.H. Freeman and CO., New York.

Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. Agricultural Systems, 89(2-3), 225-247. https://doi.org/10.1016/j.agsy.2005.11.004

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

Upton, W. H., Donoghue, K. A., Graser, H. U., & Johnston, D. J. (1999). Ultrasound proficiency testing. Meeting of the Association of Advancement in Animal Breeding and Genetics, Armidale, New South Wales, Australia. pp. 341-344. http://www.aaabg.org/proceedings/1999/AB99079.pdf

Walmsley, B. J., McPhee, M. J., & Oddy, V. H. (2014). Development of the BeefSpecs fat calculator to assist decision making to increase compliance rates with beef carcass specifications. Animal Production Science, 54(11-12), 2003-2010. https://doi.org/10.1071/An14611

Walmsley, B. J., Oddy, V. H., McPhee, M. J., Mayer, D. G., & Mckiernan, W. A. (2011). BeefSpecs a tool for the future: On-farm drafting and optimising feedlot profitability. Australian Farm Business Management Journal, 7(2), 29-36. https://doi.org/10.22004/ag.econ.121460