Predictive performance of penalized beta regression model for continuous bounded outcomes

Emmanuel O. Ogundimu

Department of Mathematics, Physics and Electrical Engineering, Northumbria University, UK

Gary S. Collins

Centre for Statistics in Medicine, University of Oxford, UK

Abstract

Prediction models for continuous bounded outcomes are often developed by fitting ordinary least square regression. However, predicted values from such method may lie outside the range of the outcome as it is bounded within a fixed range, with non-linear expectation due to the ceiling and floor effects of the bounds. Thus, regular regression models such as normal linear or nonlinear models, are inadequate for prediction purposes for bounded response variable and the use of distributions that can model different shapes are essential. Beta regression, apart from modeling different shapes and constraining predictions to an admissible range, has been shown to be superior to alternative methods for data fitting but not for prediction purposes. We take data structures into account and compared various penalized beta regression method on predictive accuracy for bounded outcome variables using optimism corrected measures. Contrary to results obtained under many regression contexts, the classical maximum likelihood method produced good predictive accuracy in terms of R^2 and RMSE. The ridge penalized beta regression performed better in terms of g-index, which is a measure of performance of the methods in external data sets. We restricted attention to prespecified models throughout and as such variable selection methods are not evaluated.

Key Words: Beta regression; Bounded outcome; Ridge regression; Prediction model; Shrinkage methods; Internal validation.

1 Introduction

Continuous bounded outcome data that take values in a finite range are common in studies of quality of life (QoL). Outcomes typically display heteroscedasticity, where the variance is smaller near the extremes, and asymmetry. This is because the distribution of the outcome is often unimodal, $U-$ shaped, and $J-$ shaped. These outcomes are seldom analysed using models that correctly constrain inference and predictions to lie within the feasible range of values (Bottai et al., 2010). The standard practice is to transform the outcome values and then carry out a standard linear regression analysis. A common transformation is to use the logit transform (also called logit normal model) that removes the boundary of the outcome and make the distribution symmetric about zero; $\tilde{y} = \log(y/1-y)$, where y is the outcome. Whilst this approach might correct the asymmetry in the data, regression models involving data from the unit interval, such as rates and proportions, are typically heteroscedastic (Cribari-Neto and Zeileis, 2010); the transformation will not stabilize the conditional variance (Kieschnick and McCullough, 2003) and the regression parameters are no longer expressed in terms of the mean of the original outcome y. An alternative to transformation is to use quantile regression. This method has the advantage that no assumption regarding the outcome distribution is made and it constrains inference within the bounded range (Bottai et al., 2010). However, it is less appealing because of less elegant interpretation than parametric models.

The use of skew-normal and censored skew-normal models have been considered for modelling bounded outcomes (Hutton and Stanghellini, 2011). Although bounded outcomes might exhibit inherent skewness, often the skewness are spurious and due to the ceiling and floor effects at the bounds. One strategy is to use models with support restricted to a fixed interval. An example of such model was considered in Flecher et al. (2010), where a truncated skew-normal model was used to analyse daily relative humidity data. The major shortcoming of using these models for predictions in bounded outcomes is that data are not observed on the restricted interval because they are not defined outside the intervals but not because they are censored or truncated. Consequently, it is more realistic to use models that have finite range and are naturally skewed for inference and predictions in bounded outcomes.

Beta distribution assumes values on the standard unit interval, and it is flexible sufficiently to model unimodal and bimodal data that are symmetric or skewed. It has been proposed and judged suitable for modelling bounded outcomes (Kieschnick and McCullough, 2003; Ferrari and Cribari-Neto, 2004). The model uses a parametrization of the beta distribution in terms of its mean and a precision parameter. In this case, the regression parameters are interpretable in terms of the original variable of interest. If the variable takes on values in (a, b) (with $a < b$ known), then the response can be transformed to the (0,1) interval by $(y-a)/(b-a)$, where b and a are the maximum and minimum possible scores. The classical and extended beta regression models have been distinguished (Cribari-Neto and Zeileis, 2010), whereby the former are similar to generalized linear models, and the precision is a constant and possibly a nuisance parameter, while the latter requires a dispersion parameter that depends on a set of predictors through a link function.

The aim of this article is to compare the classical beta regression model and its penalized version for prediction purposes in the presence of challenging data structures. These structures include multicollinearity among predictors, binary predictors with few events, noise variables and small sample sizes, which are frequently encountered in practice. We evaluated bootstrap shrinkage factor method, ridge regression, and bias correction and reduction methods on predictive accuracy of the models. Four predictive accuracy measures- R^2, RMSE, mean absolute error and Gini index- which are corrected for possible overfitting was used to assess model performance.

We restrict attention to prespecified models and as such penalized methods with variable selection are not evaluated. In section 2, we describe the real data used in the study. Shrinkage and penalized beta regression methods are discussed in section 3. The methods are compared using simulation and the data example in section 4 and conclusions are given in section 5.

2 Data description and results

The data set is from a multi-center randomized controlled trial of treatments for Whiplash Associated Disorder, the Managing Injuries of the Neck Trial (MINT), in which two treatment regimes were compared: physiotherapy versus reinforcement of advice for patients with continuing symptoms three weeks after their initial visit to the Emergency Department (ED) (Lamb et al., 2007). As with many patient-reported outcome or quality of life studies, data were collected using questionnaires at regular intervals, 4, 8 and 12 months, after patients' ED attendance (see Figure 1 in Appendix A).

The primary outcome of interest is return to normal function after the whiplash injury, and is measured using the Neck Disability Index (NDI). The NDI is a self-completed questionnaire which assess pain-related activity restrictions in 10 areas including personal care, lifting, sleeping, driving, concentration, reading and work, leading to a score between 0 and 50. Due to its reliability and validity, the NDI questionnaire is a standard instrument for measuring self-rated disability resulting from neck injuries.

Data are available from 599 patients, aged 18 to 78 years, contributing 1934 measurements over the study period, with 372 (62%) patients having complete observations (i.e. scores at all measurement occasions). For simplicity, we focus our analysis on the 372 patients with complete data. Scores at month 8 are used for model development whilst scores at month 12 are used for temporal validation.

3 Methods of estimation

3.1 Beta-regression and Maximum likelihood estimator (MLE)

Beta regression provides a framework for modelling continuous variables constrained in the standard unit interval (Ferrari and Cribari-Neto, 2004). The probability density function (*PDF*) of a beta distributed random variable y with shape parameters p and $q > 0$ is given by

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad y \in (0,1), \tag{1}$$

where $\Gamma(.)$ is the gamma function. The mean and variance y are respectively,

$$E(y) = \frac{p}{p+q} \quad \text{and} \quad Var(y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

A reparametrized version of (1) that is convenient for modelling purposes is given as

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad y \in (0,1), \tag{2}$$

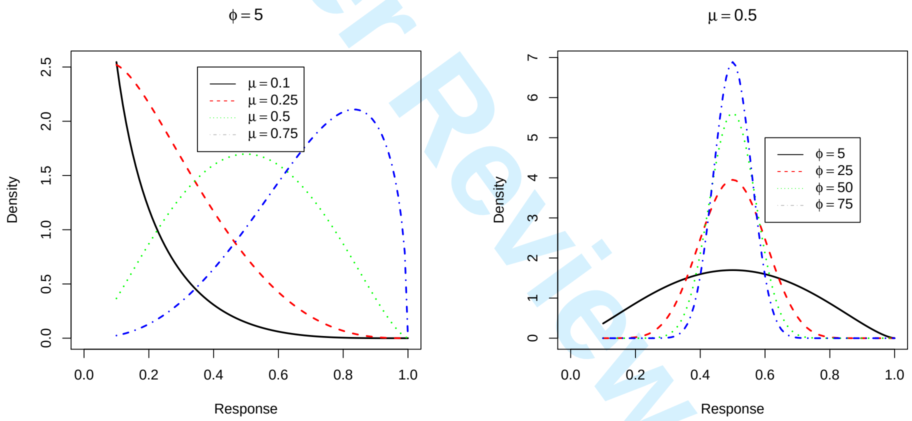
where $\mu = p/(p+q)$ and $\phi = p+q$ (equivalently $p = \mu\phi$ & $q = (1-\mu)\phi$). This parametrization implies $y \sim B(\mu,\phi)$, and thus $E(y) = \mu$ and Var$(y) = \mu(1-\mu)/(1+\phi)$. The mean is therefore μ and ϕ is the precision parameter since it controls the variance of the distribution. As stated earlier, the classical beta regression model is similar to the generalized linear model with mean parameter μ expressed as a function of covariates, and ϕ treated as nuisance parameter. A strictly monotone and twice differentiable *link function* that maps the unit interval to the real line is used to ensure that the fitted values are within the admissible range. All the link functions used in the generalized linear model can be used for this purpose. Our focus will be on the classical beta regression with logit link.

The maximum likelihood estimator is based on the log-likelihood function corresponding to the *PDF* in equation (2), and is given by $\sum_{i=1}^{n} l_i(\mu_i, \phi)$, where

$$l_i(\mu_i,\phi) = \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)\log y_i + \{(1-\mu_i)\phi - 1\}\log(1-y_i), \quad (3)$$

with $\mu \in (0,1)$ and $\phi > 0$, since p, $q > 0$. The mean is linked to the linear predictor by $g(\mu_i) = \beta' X_i$, where X_i are the predictors, β is the vector of regression coefficients and g is the logit link.

Figure 1 shows the PDFs of beta distribution with varying mean and dispersion parameters. Both figures show the flexibility of the model to assume different shapes when different values of the parameters are combined. Figure 1(b) further justifies ϕ as a dispersion parameter since the spread in the response decreases as ϕ increases.



(a) (b)

Figure 1: Probability density functions for beta distribution : (a) with varying μ = 0.10, 0.25, 0.50, 0.75 and fixed $\phi = 5$; (b) with varying ϕ = 5, 25, 50, 75 and fixed $\mu = 0.5$.

3.2 Bootstrap shrinkage factor

It has been shown that overfitting in prediction models can be reduced by applying a heuristic shrinkage factor (Moons et al., 2004). In this setting, regression coefficients obtained by maximum likelihood can be shrunk by multiplying their values by a linear shrinkage factor. This can be obtained using the nonparametric bootstrap method. The method can be described for the beta regression model as follows; a bootstrap sample is drawn from the original data set and the model is fitted using maximum likelihood. The linear predictor values are then fitted as a single predictor in a beta regression model in the original data set using maximum likelihood. This is repeated multiple times and the slopes of the regression models are obtained. The slopes are then averaged to produce a shrinkage factor. The method is referred to as a uniform or linear shrinkage method because the regression coefficients are simultaneously adjusted by a single shrinkage factor. This method will be evaluated in the data example only.

3.3 Ridge regression

Uniform shrinkage methods are shrinkage methods after estimation. Alternative approaches using penalized likelihood, like the ridge regression, apply nonuniform shrinkage factors to the maximum likelihood estimates. Suppose $l(\beta)$ is the reparametrized version of (3) in terms of β, then the ridge regression is the constrained likelihood problem where $l(\beta)$ is maximized subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$; $t \geq 0$, t is a scalar chosen by the investigator. Equivalently, the ridge penalized likelihood function can be written as

$$l(\beta^{rd}) = l(\beta) - \lambda \sum_{j=1}^{p} \beta_j^2, \tag{4}$$

where $\lambda \geq 0$ has a one-to-one correspondence with t, and it is referred to as the regularization parameter and β^{rd} are the parameter estimates from the ridge regression. The larger the value of λ, the further the parameter estimates are shrunk towards zero. In particular, the standard beta model is recovered when $\lambda = 0$ in (4). The ridge estimator is a convex optimization problem and it is sometimes referred to as an L_2-type regularization procedure.

3.4 Bias correction and reduction

The method of ridge regression introduces bias in the maximum likelihood estimator in order to gain lower variance. Other methods have been proposed in the literature that correct or reduce bias in finite samples. The extent to which these correctional approaches improve prediction is uncertain. Suppose μ is the true parameter and $\hat{\mu}$ is its estimates, the bias of the estimator of μ is defined as

$$B(\mu) = E_{\mu}(\hat{\mu} - \mu),$$

where E_{μ} is expectation taken over μ. We distinguish two methods for eliminating this bias.

3.4.1 Bias Correction

Simulation based approaches (Jackknife and Bootstrap) have been discussed in the literature for bias correction (Kosmidis, 2014). We focus on the method based on the asymptotic bias correction of the MLE, $\hat{\mu}$, which can be expressed as

$$B(\mu) = \frac{b_1(\mu)}{n} + \frac{b_2(\mu)}{n^2} + \dots.$$

Methods that correct higher order bias $b_1(\mu)/n + b_2(\mu)/n^2$ have been discussed elsewhere (see Cordeiro and Barroso (2007)). We focus on the method that removes the first-order bias term for beta regression, given by

$$\tilde{\mu} = \hat{\mu} - \frac{b_1(\hat{\mu})}{n} \text{ and } E_\mu(\tilde{\mu} - \mu) = O(n^{-2}). \tag{5}$$

A closed form expression for (5) is given in Cordeiro and Klein (1994). Details of the expression for $b_1(\mu)/n$ for beta regression models can be found in Grun et al. (2012). The main drawback of this method (and other bias correction methods, such as jacknife and bootstrap) is that it depends on the finiteness of $\hat{\mu}$ and also inherits all the instability in its estimator.

3.4.2 Firth's modified likelihood

Firth (1993) introduced a penalized maximum likelihood estimation method for the reduction of bias in MLEs. The method has been adapted for solving monotone likelihood problems in other regression contexts (Heinze and Schempe, 2001; Heinze and Schemper, 2002). The beta regression likelihood function can be penalized using this approach, and is given as

$$l(\beta^\star) = l(\beta) + \frac{1}{2}\log|I(\beta)|. \tag{6}$$

The penalty function $|I(\beta)|^{1/2}$ is known as Jeffreys invariant prior. The superiority of Firth method over other post hoc bias correction of MLEs is that the estimates of the penalized likelihood almost always exist. As ridge estimator is guaranteed to reduce variance, it is possible for bias correction or reduction methods to inflate the estimator's variance. A comparative study of bias correction and reduction methods in logistic regression has been considered in Maiti and Pradham (2007), but not with a view towards prediction.

3.5 Software for beta regression

Many statistical packages have routine for fitting the classical beta regression. *Betareg* package in R (Cribari-Neto and Zeileis, 2010) and in Stata used MLE. Non-linear routines in SAS such as PROC NLIN or NLMIXED can also be used. Bayesian approach for beta regression has been implemented in the *Bayesianbetareg* in R. We used *betareg* in R for the classical maximum likelihood estimates, bias corrected and bias reduced beta regression. For ridge regression, we used the function *ri()* in GAMLSS package in R, which automatically estimates shrinkage parameter (and therefore the effective degrees of freedom) using a local maximum likelihood procedure (Rigby and Stasinopoulos, 2013).

4 Simulation and data analysis

4.1 Simulation studies

To evaluate the performance of ridge penalized beta regression for prediction in bounded outcomes, we compared it with bias reduction, bias correction and classical beta regression methods. We simulated 7 variables- two binary predictors and 5 continuous predictors. To begin with, we fixed the regression parameters at 0.5, correlation between variable at zero and prevalence of binary predictors at 0.5. The following scenarios were investigated.

1. Impact of sample size: We simulated 25, 50, 100, 200, and 500 observations. This corresponds to 3.6, 7.1, 14.2, 28.6 and 71.4 subjects per variable (SPV) with 7 variables respectively.

2. Impact of noise variables: Noise variables are variables whose true regression coefficients are zero. These variables are common in prespecified models. We added 2 binary and 3 continuous predictors not related to the outcome variable to scenario 1. This significantly reduced the SPV of the models across the sample sizes.

3. Impact of low prevalence binary predictors: This implies binary predictors with rare events e.g. the use of binary variable denoting presence/absence of penile cancer, known to be rare, for the prediction of cardiovascular diseases in men. We varied the prevalence of the simulated binary predictors between 0.03 and 0.5 for the sample size of 500.

4. Impact of correlation: We varied the correlation between two of the continuous predictors. The correlations considered include 0, 0.1, 0.3, 0.5, 0.7 and a perfect correlation where one of the variable is a linear combination of the other.

Simulation studies comparing predictive performance of statistical method usually consider apparent performance as a means of method evaluation. We take a different approach and report optimism corrected measures instead. This is due in part to potential overfitting of the models especially in small samples. Cross-validation and bootstrap methods are common approaches for optimism quantification. We focus on the use of bootstrap approach, which is based on the algorithm described below.

- Simulate a data set based on the appropriate data structure and fit beta model or its penalized version to the data. *Apparent* predictive performance based on the measure of interest, say R^2_{app} is computed on this model

- From the simulated data set, take say, M bootstrap sample, where $m = 1 \ldots M$ and fit beta model or its penalized version and compute $R^{2(m)}_{boot}$

- For each of the M models from the bootstrap samples, compute R^2 back in the original data, say $R^{2(m)}_{orig}$

- Optimism in model fit for each of the bootstrap sample is $Opt^m = R^{2(m)}_{boot} - R^{2(m)}_{orig}$. The average optimism is

$$Opt = \frac{\sum_{m=1}^{M} Opt^{(m)}}{M}$$

- Optimism adjusted measures for the original model is $R^2_{adj} = R^2_{app} - Opt$.

The predictive accuracy measures considered include pseudo R^2, RMSE, Gini's mean index (g-index) and mean absolute deviation. Briefly, R^2 indicates the proportion of information in the data that is explained by the model while RMSE measures how far, on average, the residuals (observed scores - fitted scores) are from zero. The g-index is a robust and highly efficient measure of variability in the predicted scores. It is the mean absolute difference between any two distinct elements of the vector containing the predicted scores from the fitted models. The lower the g-index, the lower the variability of the predicted scores, and the better the performance of the method in an external validation data set.

Figure 2 shows the impact of sample size on the predictive accuracy measures when classical maximum likelihood (ML), bias corrected ML (BC), bias reduced ML (BR) and ridge penalized beta regression are compared. As expected the performance of the models improved as sample size increases. ML, BC and BR are indistinguishable on R^2, RMSE and mean absolute deviation as predictive accuracy measures. However, the BC and BR methods showed slight improvement over ML up to sample size of 200 as the predicted scores are less variable (g-index). This signifies potential improved predictions in external validation data sets. The ridge penalized method uniformly outperformed other methods across the sample sizes using this measure.

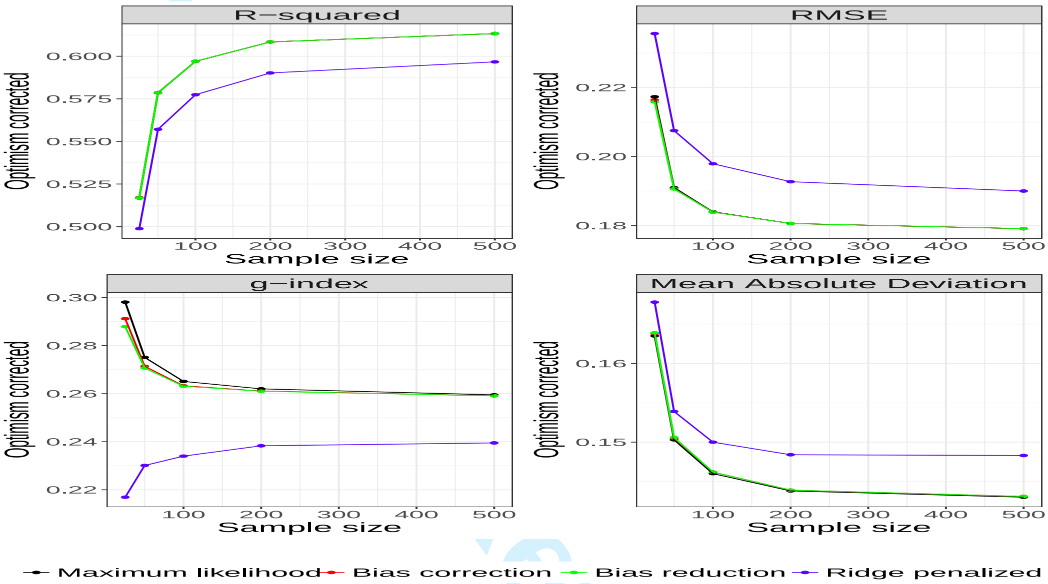


Figure 2: Optimism corrected measures showing impact of sample size ({25, 50, 100, 200, 500} on predictive accuracy

The results of predictive accuracy measures when 5 noise variables are added to the model is shown in Figure 3. The conclusions are similar to what we observed in Figure 2 except that the performance of methods in Figure 2 are much better. This result is intuitive and has been evaluated in similar studies with binary and time-to-event data. The implication of this is that the so called subjects per variable, SPV (equivalently events per variable for binary and time-to-event data) reduced dramatically and thus have significant effects on the fit and model predictive accuracy.

The assessment of the impact of prevalence on predictive accuracy of the methods show that the performance of ML, BC and BR methods are similar (see Figure 2 in Appendix B). These methods outperformed the ridge method in terms of R^2, RMSE and mean absolute deviation. The ridge regression performs better in terms of lower variability in the predicted scores (g-index). The R^2 suggests that prevalence of 0.3 is optimal and that there is no added merit of using binary predictors with prevalence of 0.5.

The results of the impact of correlation show that predictive accuracy improves as correlation increases. The performance of ML, BC and BR methods are again similar. As expected, the ridge regression is the only method that yields finite parameter estimates in the presence of perfect correlation. The results also indicate that the higher the correlation between the predictors the lower the prediction accuracy in an external validation sample (see Figure 3 in

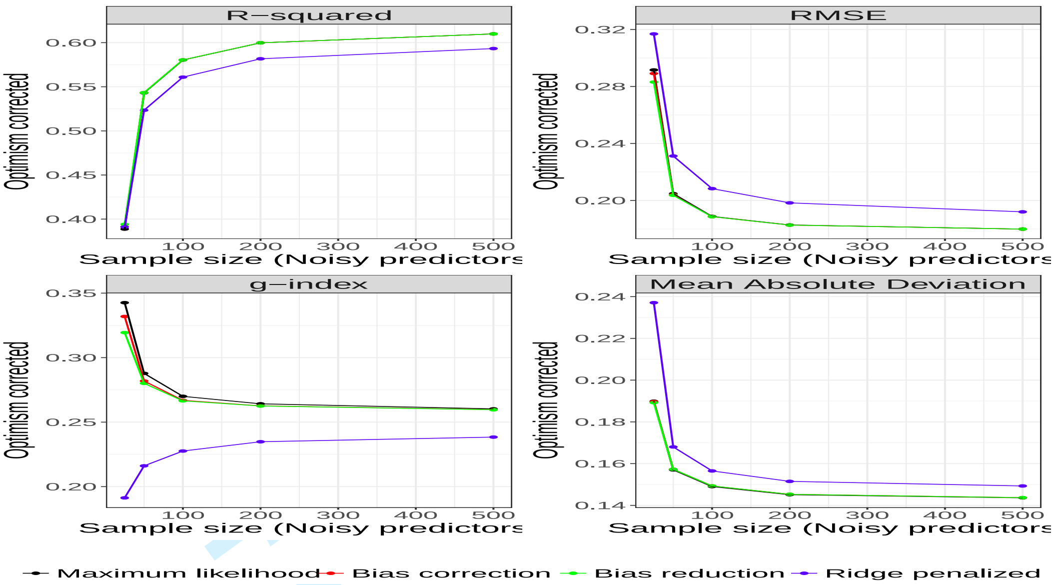


Figure 3: Optimism corrected measures showing impact of 5 noise variables added to the 7 main variables on predictive accuracy

Appendix B).

The results of the comparison of apparent and optimism corrected measures indicate that the apparent ML is slightly higher than the apparent ridge regression values for sample size smaller than 100 observations. The comparison of apparent and optimism corrected absolute mean deviations for the ridge regression showed that the former exaggerated the prediction errors across the sample sizes, whereas its adjusted values are in agreement with the ML corrected values (see Figure 4 in Appendix B).

4.2 Data analysis

We used the methods described in section 3 to develop a prediction model for the NDI data set. The model was developed using NDI scores at the third measurement occasion (Month 8). We evaluated its predictive performance using the data at the last measurement occasion (Month 12) - the so-called temporal validation. Instead of the RMSE evaluated in the model development data we used root mean square prediction error, RMSPR as its external validation equivalent. It is given as $\sqrt{\frac{1}{n^*}\sum(Y_i - \hat{Y}_i)^2}$, where $n^\star$ is the sample size at month 12, $\hat{Y}$ is the predicted values using regression coefficients from the model developed and Y_i is the observed response at month 12. Similarly, R^2 was replaced by R_p^2 as its external validation equivalent, and it is given as $1 - \sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2 \Big/ \sum_{i=1}^{n^*}(Y_i - \bar{Y})^2$, where $\bar{Y}$ is the arithmetic mean of the dependent variable in the data set used for model development.

The data set contains 4 predictors that are related to the outcome. These predictors have been used previously in other context (see Ogundimu and Hutton (2015)). Initial analysis showed that there is no difference in the performance of the methods in terms of predictive accuracy. In particular, the uniform shrinkage value is 1.0001. This is not different from the ML method. We therefore added 10 noise variables to the model to evaluate their impact. This reduced the uniform shrinkage value to 0.9701 (see Table 1 in Appendix A for the internal validation results).

Table 1 shows the performance of the methods in predicting NDI scores at month 12. The ridge regression shows a slightly better predictive performance than the ML, BC and BR methods.

Table 1: Model performance in predicting NDI scores at Month 12

	R_p^2	RMSPR	g-index	Mean absolute deviation
Bootstrap shrinkage	0.4900	0.1441	0.1437	0.1197
Maximum likelihood	0.4899	0.1423	0.1464	0.1175
Bias correction	0.4898	0.1425	0.1460	0.1177
Bias reduction	0.4898	0.1425	0.1460	0.1176
Ridge penalized	0.4910	0.1181	0.1304	0.0932

5 Concluding Remarks

Continuous bounded outcome data are frequently encountered in mental health, social sciences and political science research and there is increasing interest on how to use such data for prediction purposes. This paper evaluated the predictive performance of penalized beta regression for continuous bounded outcome. Unlike the maximum likelihood estimator (MLE) of the logistic model, the MLE for beta model had less difficulties with model convergence in the presence of challenging data structures, such as binary predictors with rare events and multicollinearity. We have shown, through a simulation study, that the predictive performance of the maximum likelihood (ML) method is not worse than the bias correction (BC) and Bias reduction (BR) methods. Theoretically, the BC and BR methods are not optimized for predictions. So, the notion that the BC and BR methods will yield better predictions due to lower bias than the ML method is not true. There is no guarantee that the methods will not result in inflated variance. However, the ridge penalized beta regression showed better predictive accuracy in external validation sample, and it is recommend for predictions over the ML, BC and BR methods. The justification for this is the bias-variance trade off of the ridge regression, which has been observed in other regression context such as the logistic regression and Cox model.

Although we focus here on prespecified models, various method for improved predictions in other regression contexts can be easily adapted to beta regression model. It would be interesting to explore, for example, the predictive performance of elastic net penalty which incorporates both the LASSO (least absolute shrinkage and selection operator) for variable selection and the ridge penalty to optimize predictions for beta regression models.

Acknowledgements

The authors are grateful to Professor S Lamb and the MINT team for access to the data.

References

Bottai, M., B. Cai, and R. McKeown (2010). Logistic quantile regression for bounded outcomes. *Stat Med* **29**, 309–317.

Cordeiro, G. and L. Barroso (2007). A third-order bias corrected estimate in generalized linear models. *Test* **16**, 76–89.

Cordeiro, G. and R. Klein (1994). Bias correction in arma models. *Stat Probabil Lett* **19**, 169–176.

Cribari-Neto, F. and A. Zeileis (2010). Beta Regression in R. *J Stat Softw* **34(2)**, 1–24.

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *J Appl Stat* **31(7)**, 799–815.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Flecher, C., D. Allard, and P. Naveau (2010). Truncated skew-normal distributions: moments, estimation by weighted moments and application to climate data. *Metron* **68**, 331–345.

Grun, B., I. Kosmidis, and A. Zeileis (2012). Extended beta regression in r: Shaken, stirred, mixed, and partitioned. *J Stat Softw* **48(11)**, 1–25.

Heinze, G. and M. Schempe (2001). A solution to problem of monotone likelihood in cox regression. *Biometrics* **57**, 114–119.

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Stat Med* **21**, 2409–2419.

Hutton, J. and E. Stanghellini (2011). Modelling bounded health scores with censored skew-normal distributions. *Stat Med* **30**, 368–376.

Kieschnick, R. and B. McCullough (2003). Reanalysis of variates observed on (0,1): percentages, proportions and fractions. *Stat Modelling* **3**, 193–213.

Kosmidis, I. (2014). Bias in parametric estimation: reduction and useful side-effects. *WIREs Comput Stat* **6**, 185–196.

Lamb, S., S. Gates, M. Underwood, M. Cooke, D. Ashby, A. Szczepura, M. Williams, E. Williamson, M. Withers, S. Isa, and A. Gumber (2007). Managing Injuries of the Neck Trial (MINT): design of a randomised controlled trial of treatments for whiplash associated disorders. *BMC Musculoskelet Disord* **8**, 7.

Maiti, T. and V. Pradham (2007). A comparative study of the bias corrected estimates in logistic regression. *Stat Methods Med Res* **17**, 621–634.

Moons, K., A. Donders, E. Steyerberg, and F. Harrell (2004). Penalized maximum likelihood estimation to directly adjust diagnostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol* **57**, 1262–1270.

Ogundimu, E. O. and J. L. Hutton (2015). A sample selection model with skew-normal distribution. *Scand J Stat* **43**, 172–190.

Rigby, R. and D. M. Stasinopoulos (2013). Automatic smoothing parameter selection in gamlss with an application to centile estimation. *Stat Methods Med Res* **23**, 318–332.

Supplementary material to: Predictive performance of penalized beta regression model for continuous bounded outcomes

Emmanuel O. Ogundimu

Department of Mathematics, Physics and Electrical Engineering, Northumbria University, UK

Gary S. Collins

Centre for Statistics in Medicine, University of Oxford, UK

Appendix A

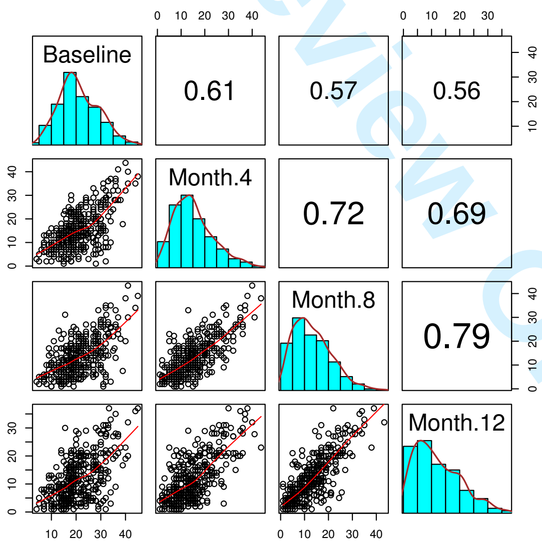


Figure 1: Scatter plots, histograms and correlations for the NDI scores at Baseline, Month 4, 8 and 12

The correlation tends to decrease with increasing time separation between the measurement times. The histograms shows that skewness is evident in the data and it becomes increasing noticeable as time progresses. The skewness could be due to selective dropout over time or the boundedness of the response variable. Hence, the use of a class of model that can model various shapes.

Table 1: Optimism corrected model performance in predicting NDI scores at Month 8

	R^2	RMSE	g-index	Mean absolute deviation
Bootstrap shrinkage	0.5530	0.1109	0.1327	0.0862
Maximum likelihood	0.5494	0.1112	0.1339	0.0860
Bias correction	0.5494	0.1112	0.1337	0.0860
Bias reduction	0.5494	0.1112	0.1337	0.0860
Ridge penalized	0.5494	0.1114	0.1301	0.0835

Table 1 shows the performance of the optimism adjusted measures using the 5 methods considered. Unlike in the case where all the variables are associated with the outcome, the uniform shrinkage value in this case is 0.9701. This shows that some form of regularization is essential to improve predictive accuracy. The uniform shrinkage method outperforms ML, BC, BR and ridge methods on R^2, and RMSE. The performance of ML BC and BR are essentially the same as we observed in the simulation study. The ridge method showed through the g-index and mean absolute deviation that it can perform better than its competitors in external validation samples.

Appendix B

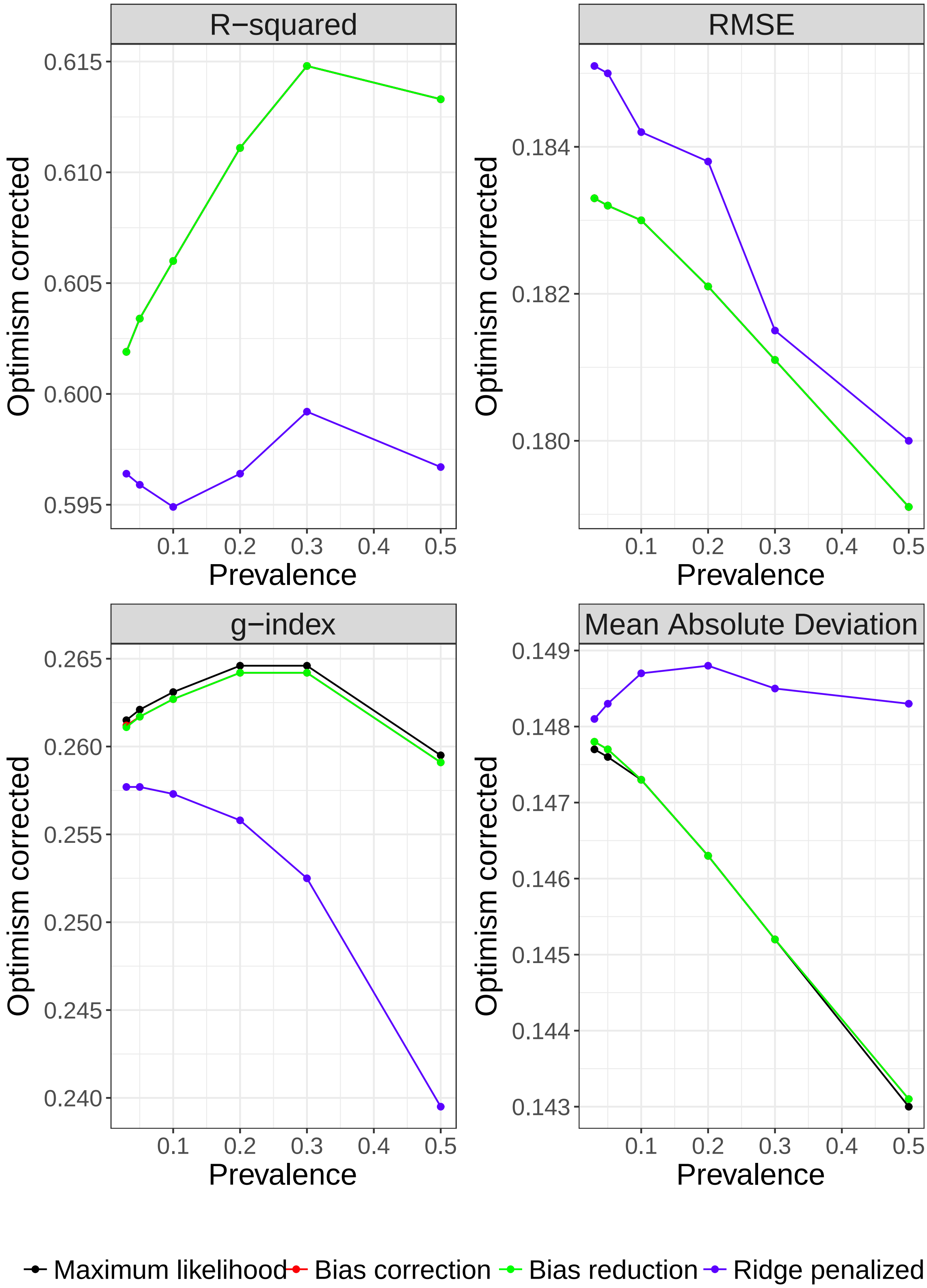


Figure 2: Optimism corrected measures showing impact of prevalence on predictive accuracy

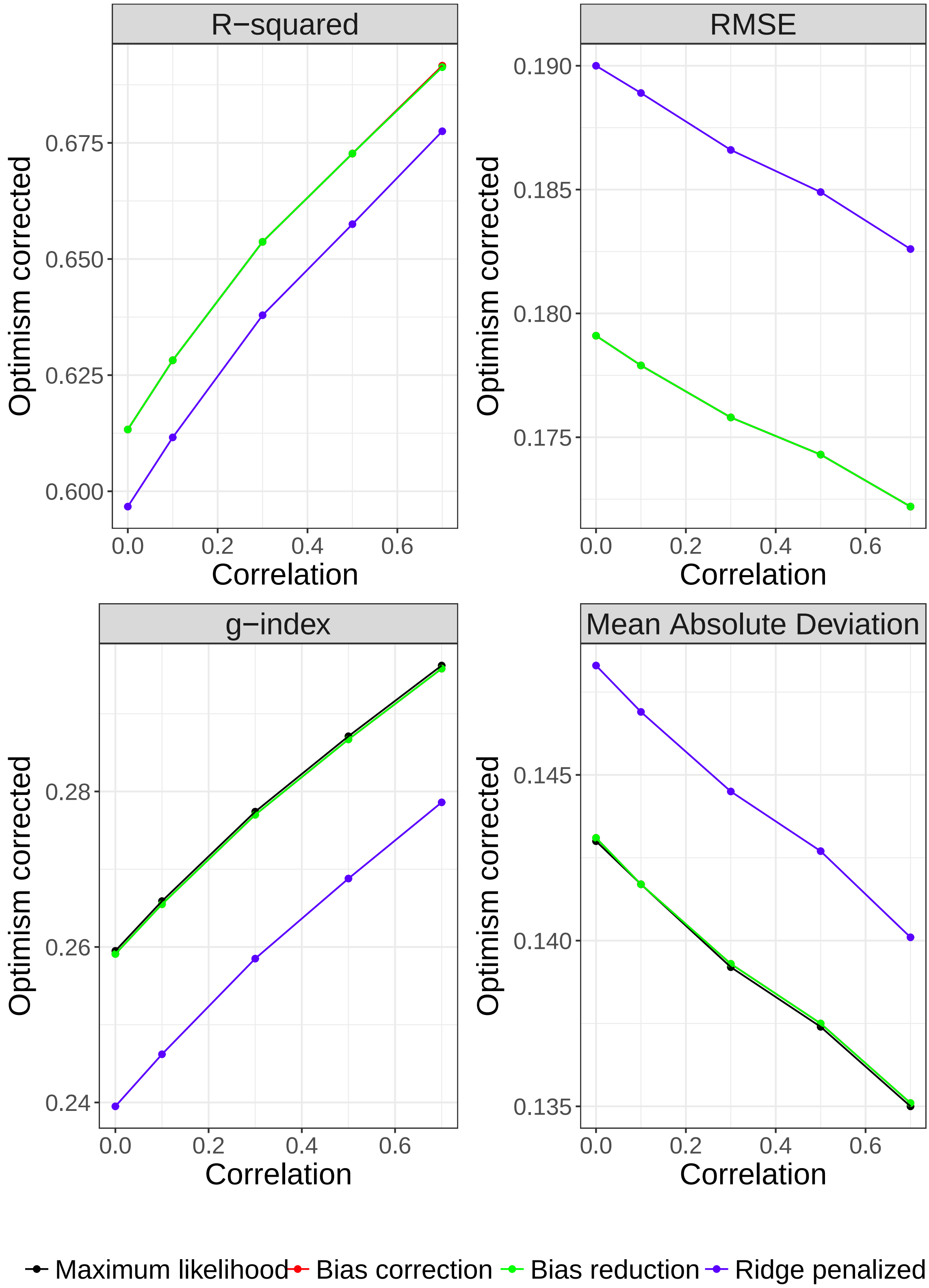


Figure 3: Optimism corrected measures showing impact of correlation on predictive performance

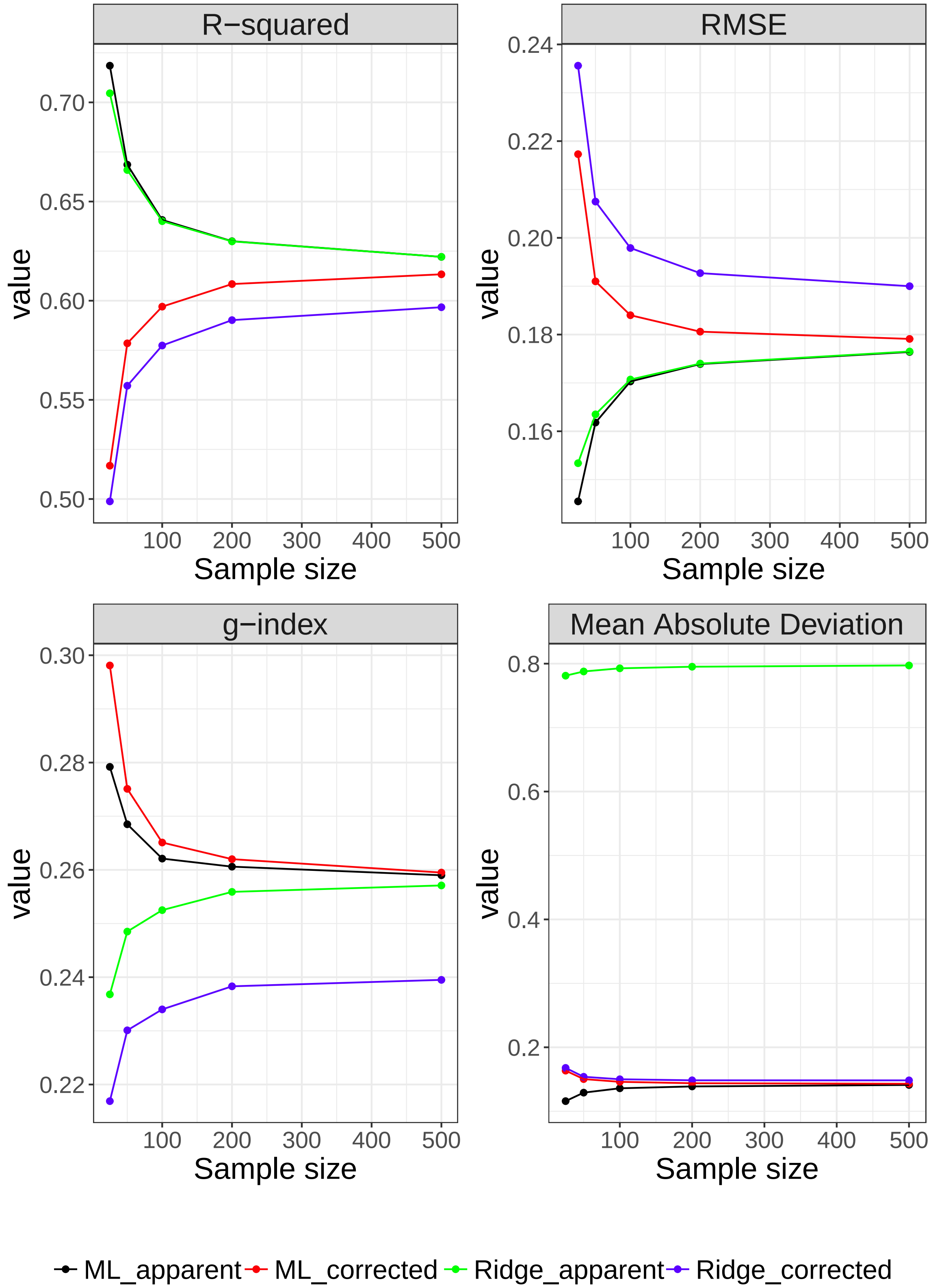


Figure 4: Apparent vs optimism adjusted measures for maximum likelihood and Ridge penalized beta regression