Running headline: Prediction error - missing data

Title: THEORETICAL EVALUATION OF PREDICTION ERROR IN LINEAR REGRESSION WITH A BIVARIATE RESPONSE VARIABLE CONTAINING MISSING DATA

Lars Erik Gangsei
Animalia - Norwegian Meat and Poultry Research Centre
P.O. 396 - Økern
N-0513 OSLO, Norway
lars.erik.gangsei@animalia.no
+47 950 61 231

Trygve Almøy
Dep. of Chemistry, Biotechnology, and Food Science
Norwegian University of Life Sciences
P.O. Box 5003
N-1432 Ås, Norway
trygve.almoy@nmbu.no

Solve Sæbø
Dep. of Chemistry, Biotechnology, and Food Science
Norwegian University of Life Sciences
P.O. Box 5003
N-1432 Ås, Norway
solve.sabo@nmbu.no

Corresponding author:

Lars Erik Gangsei
Animalia - Norwegian Meat and Poultry Research Centre

P.O. 396 - Økern

N-0513 OSLO, Norway

lars.erik.gangsei@animalia.no

+47 950 61 231

Key Words: bivariate linear regression; James Stein estimator; missing data; prediction error; risk function.

ABSTRACT

Methods for linear regression with multivariate response variables are well described in statistical literature. In this study we conduct a theoretical evaluation of the expected squared prediction error in bivariate linear regression where one of the response variables contains missing data. We make the assumption of known covariance structure for the error terms. On this basis, we evaluate three well-known estimators; standard ordinary least squares, ~~weighted~~ generalized least squares and a James-Stein inspired estimator. Theoretical risk functions are worked out for all three estimators to evaluate under which circumstances it is advantageous to take the error covariance structure into account.

1 Introduction and notation

In this paper, we evaluate a linear regression model with a bivariate response variable where one of the responses contains missing data. For practical purposes this situation is likely to occur if predictor variables and one response variable, typically a response variable of subordinate interest, are easily sampled; but the other response, typically the one of primary interest, is hard(er) or more costly to sample. Often, though not necessarily, the fully observed response variable will be a surrogate variable (Upton and Cook, 2014) for the response variable containing missing data. In such situations, a sampling method where the primary response variable is sampled only for a subset of the total sample, might be beneficial, especially if the error terms in the bivariate linear regression model are highly correlated.

When no data is missing and ordinary least squares (OLS) estimators are used, the gain of applying one single multi-response regression model is limited compared to several single response models, since the regression parameter estimates are equal and unaffected by the covariance structure of the error term. In this paper, we show that the gain of using a bivariate response variable model in cases with missing data for one of the responses might be substantial if the covariance structure of the error terms is taken properly into account.

In the following scalars are denoted by ~~lower case~~ lowercase italic characters, vectors by ~~lower case~~ lowercase bold italic characters and matrices by upper case bold italic characters. The single elements of any matrix are denoted by the corresponding ~~lower case~~ lowercase italic letter and a subindex, i.e. w_{ij} is the element of the ith row and jth column of $\boldsymbol{W}$. In general, Greek letters are used for parameters and Latin letters are used for random variables. The risk function ~~for an estimator or predictor~~, for an estimator or predictor, $\hat{\boldsymbol{\theta}}$, ~~and the~~ with true value, $\boldsymbol{\theta}$, $R_{\hat{\boldsymbol{\theta}}}$, is

$$R_{\hat{\boldsymbol{\theta}}} = E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right];$$

In the present paper we evaluate a regression model where we assume the error covariance

term to be known. In a companion paper, Gangsei et al. (2016b), estimators based on unknown error covariance structure are evaluated. The theoretical work presented in the present paper and Gangsei et al. (2016b) were initiated by a practical problem of estimating the Lean Meat Percentage in Norwegian pork carcasses. In Gangsei et al. (2016a) the methods were successfully applied to this practical problem.

2 Model spesification

The data are given by an $n_1 \times 2$ matrix of response variables, $\boldsymbol{Y}_1$, and an $n_1 \times p$ matrix of predictor variables, $\boldsymbol{X}_1$, in which the first column is the vector of unity, and the $p-1$ last columns are denoted $\boldsymbol{Z}_1$. If not stated otherwise, $\boldsymbol{Z}_1$ is assumed to be mean centred. The model is a standard bivariate response variable regression model, i.e.

$$\boldsymbol{y}_i^T \sim N_2\left(\boldsymbol{\beta}^T \boldsymbol{x}_i^T,\ \boldsymbol{\Sigma}\right),\ \ i = 1, \ldots, n_1$$

where $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ denote the ith row of $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$ respectively. The $p \times 2$ matrix $\boldsymbol{\beta}$, which first and second column are denoted $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, denotes the regression coefficients. The notations $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_z$ are used for the first and $p-1$ last rows of $\boldsymbol{\beta}$. The 2×2 matrix $\boldsymbol{\Sigma}$ denotes the error covariance matrix, with elements σ_{ij}, $i = 1, 2$, $j = 1, 2$. The model is well known from literature, also in a Bayesian setting (Box and Tiao, 1973; Minka, 2000).

We assume that the data represents a random sample from a larger population, of which a random subsample contains missing data for the second response variable. The observations are rearranged so that the first n_2 $(p < n_2 \leq n_1)$ rows of $\boldsymbol{Y}_1$ are fully observed, and for the $n_1 - n_2$ last rows of $\boldsymbol{Y}_1$ only the first column is observed. For the rest of this paper $\boldsymbol{Y}_2$, $\boldsymbol{X}_2$ and $\boldsymbol{Z}_2$ will represent the $n_2 \times 2$, $n_2 \times p$ and $n_2 \times (p-1)$ sub-matrices of the n_2 first rows of $\boldsymbol{Y}_1$, $\boldsymbol{X}_1$ and $\boldsymbol{Z}_1$, respectively. Further, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ denote the first and second column of $\boldsymbol{Y}_1$, $\boldsymbol{y}_{21}$ and $\boldsymbol{y}_{22}$ denote the first and second column of $\boldsymbol{Y}_2$ and $\boldsymbol{y}_v$ is the stacked column-vector of $\boldsymbol{y}_1$ and $\boldsymbol{y}_{22}$.

The model might be defined in different ways, but the representation

$$\boldsymbol{y}_v \sim N_{n_1+n_2}\left(\boldsymbol{X}_{(+)}\boldsymbol{\beta}_v,\ \ \boldsymbol{\Sigma}_{(+)}\right); \tag{1}$$

where $\boldsymbol{\beta}_v$ is the stacked column-vector of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, is suitable for the purpose of the rest of this paper. The $(n_1+n_2)\times(n_1+n_2)$ covariance matrix $\boldsymbol{\Sigma}_{(+)}$ is the upper left block of $\boldsymbol{\Sigma}\otimes\boldsymbol{I}_{n_1}$. Likewise $\boldsymbol{X}_{(+)}$ denotes the $(n_1+n_2)\times 2p$ matrix representing the n_1+n_2 first rows of $\boldsymbol{I}_2\otimes\boldsymbol{X}_1$.

3 Estimators

3.1 The ~~weighted~~ generalized least squares estimator

The $2p\times 1$ vector $\hat{\boldsymbol{\beta}}_v$ denotes the ~~weighted~~ generalized least squares (GLS) estimator based on (1). We denote the first p elements of $\hat{\boldsymbol{\beta}}_v$ by $\hat{\boldsymbol{\beta}}_{v1}$, and the last p elements by $\hat{\boldsymbol{\beta}}_{v2}$. The expression and distribution of $\hat{\boldsymbol{\beta}}_v$ are:

$$\hat{\boldsymbol{\beta}}_v = \left(\boldsymbol{X}_{(+)}^T\boldsymbol{\Sigma}_{(+)}^{-1}\boldsymbol{X}_{(+)}\right)^{-1}\boldsymbol{X}_{(+)}^T\boldsymbol{\Sigma}_{(+)}^{-1}\boldsymbol{y}_v, \quad \hat{\boldsymbol{\beta}}_v \sim N_{2p}\left[\boldsymbol{\beta}_v,\ \ \left(\boldsymbol{X}_{(+)}^T\boldsymbol{\Sigma}_{(+)}^{-1}\boldsymbol{X}_{(+)}\right)^{-1}\right];$$

Remark 1

$\hat{\boldsymbol{\beta}}_{v1}$, equals $\hat{\boldsymbol{\beta}}_{11}$, i.e. the ordinary OLS estimator based on $\boldsymbol{X}_1$.

Remark 2

The expression and distribution for the ~~weighted least squares~~ GLS estimator for $\boldsymbol{\beta}_2$ is:

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_{v2} = \hat{\boldsymbol{\beta}}_{22} + \sigma_{12}/\sigma_{22}\left(\hat{\boldsymbol{\beta}}_{11} - \hat{\boldsymbol{\beta}}_{21}\right),\\
&\hat{\boldsymbol{\beta}}_{v2} \sim N_p\left\{\boldsymbol{\beta}_2,\ \ \sigma_{22}\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1} - {\sigma_{12}}^2/\sigma_{22}\left[\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\right]\right\};
\end{aligned} \tag{2}$$

, where $\hat{\boldsymbol{\beta}}_{21}$ and $\hat{\boldsymbol{\beta}}_{22}$ are the standard OLS estimates based on the n_2 first (full) observations for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively. We observe that as n_1 increases towards infinity the distribution for $\hat{\boldsymbol{\beta}}_{v2}$ approaches the distribution of $\hat{\boldsymbol{\beta}}_{22}$ conditional on known $\boldsymbol{\beta}_1$, i.e.

$$\hat{\boldsymbol{\beta}}_{22}\mid\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_{22} + \sigma_{12}/\sigma_{11}\left(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_{21}\right), \quad \hat{\boldsymbol{\beta}}_{22}\mid\boldsymbol{\beta}_1 \sim N_p\left[\boldsymbol{\beta}_2,\ \ \left(\sigma_{22} - \sigma_{12}^2/\sigma_{11}\right)\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1}\right];$$

Remark 3

The covariance between $\hat{\boldsymbol{\beta}}_{v1}$ and $\hat{\boldsymbol{\beta}}_{v2}$ is $\sigma_{12}\left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}$.

3.2 James-Stein estimator

An alternative estimator for $\boldsymbol{\beta}_v$, $k\hat{\boldsymbol{\beta}}_v$, where $0 \leq k \leq 1$ is known as the James–Stein estimator (James and Stein, 1961; Efron and Morris, 1973). It is well known that if the regularization parameter, k, is set appropriately, then the James Stein estimator outperforms the least squares estimator in the sense of having smaller risk, i.e. $R_{k\hat{\boldsymbol{\beta}}v} \leq R_{\hat{\boldsymbol{\beta}}v}$.

An obvious objection to the James Stein estimator is that it is biased for all $k \neq 1$. Another issue is how to set the regularization parameter at a suitable value. Bock (1975) showed how to set k to minimize $R_{k\hat{\boldsymbol{\beta}}v}$, based on the largest eigenvalue of covariance for $\hat{\boldsymbol{\beta}}_v$. A problem with the estimator $k\hat{\boldsymbol{\beta}}_v$ is that the same regularization, k, is applied to both $\hat{\boldsymbol{\beta}}_{v1}$ and $\hat{\boldsymbol{\beta}}_{v2}$. Brown and Zidek (1980) and Matsuda and Komaki (2015), addresses this problem by analysing different regularization matrices in detail, also based on known error covariance structure, however, they do not deal with missing data for the response variables.

In this paper, we evaluate a variant of the James–Stein estimator, $\tilde{\boldsymbol{\beta}}$, a stacked vector of $\tilde{\boldsymbol{\beta}}_1 = k_1\hat{\boldsymbol{\beta}}_{v1}$ and $\tilde{\boldsymbol{\beta}}_2 = k_2\hat{\boldsymbol{\beta}}_{v2}$ where $0 \leq k_i \leq 1$ for $i = 1, 2$.

4 Prediction error

4.1 General form: No assumptions on predictor variables

Let $\boldsymbol{y}_N$ denote a new observation, and let $\tilde{\boldsymbol{y}}_N$ denote the corresponding predicted value based on the new predictor variable, $\boldsymbol{x}_N$, i.e. a vector of length p where the first element is 1 and the last $p-1$ elements are denoted $\boldsymbol{z}_N$.

Theorem 1: Expected prediction error

The expected squared prediction errors denoted R_{yNi}, for $i = 1, 2$, using the estimator $\tilde{\boldsymbol{\beta}}_i$, and the notation $\boldsymbol{Z}_{2c}$ for the first n_2 rows of $\boldsymbol{Z}_1$ centred with respect to the column means of

the same rows, are:

$$
\begin{aligned}
R_{y_{N1}} =& \sigma_{11} + \sigma_{11} k_1^2 \left[1/n_1 + \boldsymbol{z}_N^T \left(\boldsymbol{Z}_1^T \boldsymbol{Z}_1 \right)^{-1} \boldsymbol{z}_N \right] + (1-k_1)^2 \boldsymbol{x}_N^T \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^T \boldsymbol{x}_N, \\
R_{y_{N2}} =& \sigma_{22} + \left(\sigma_{22} - \sigma_{12}^2/\sigma_{11} \right) k_2^2 \left[1/n_2 + \boldsymbol{z}_N^T \left(\boldsymbol{Z}_{2c}^T \boldsymbol{Z}_{2c} \right)^{-1} \boldsymbol{z}_N \right] + \\
& \sigma_{12}^2/\sigma_{11} k_2^2 \left[1/n_1 + \boldsymbol{z}_N^T \left(\boldsymbol{Z}_1^T \boldsymbol{Z}_1 \right)^{-1} \boldsymbol{z}_N \right] + (1-k_2)^2 \boldsymbol{x}_N^T \boldsymbol{\beta}_2 \boldsymbol{\beta}_2^T \boldsymbol{x}_N;
\end{aligned} \tag{3}
$$

The proof is deferred to the Appendix.

4.2 Multivariate normal distributed predictor variables

The formulas given by (3) is valid for a new and observed observation $\boldsymbol{x}_N$ and the given calibration set $\boldsymbol{X}_1$. A more general statement about prediction error is the expected squared prediction error over all calibration samples and new observations, ~~$E\left(R_{y_{Ni}}\right)$~~$E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right)$, which can be obtained under certain assumptions (Helland and Almøy, 1994).

Assume that all rows of the original (non-centred) $\boldsymbol{Z}_1$, and the new (non-centred) observation, $\boldsymbol{z}_N$, are independent multivariate normal distributed with fixed expectation parameter and a fixed covariance matrix, $\boldsymbol{\Gamma}$. Under these assumptions the expected prediction risks might be given as functions of $\boldsymbol{\Sigma}$, n_i, k_i, β_{0i} and R_i^2, where R_i^2 ~~might be given the interpretation as~~ is the population coefficients of determination.

The natural choices for k_i, denoted $k_{Oracle\ i}$, for $i = 1, 2$, are the values that minimize the expected squared prediction error. The sub-indexing "Oracle" is used in line with Wasserman (2006) and reflects that these values are unattainable in most practical situations.

To compare the precision of different estimators we use the expected ratios of the expected squared prediction errors of the estimators. Since the ratio of expectation is constant, this ratio is equal to ~~$E\left(R_{\tilde{y}_{Ni}}\right)/E\left(R_{\hat{y}_{Ni}}\right)$~~ $E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right)/E_{\boldsymbol{x}}\left(R_{\hat{y}_{Ni}}\right)$, where the subscripts $\tilde{y}_{Ni}$ and $\hat{y}_{Ni}$ indicate the estimator.

The expressions given in (4) simplify equations and increase readability. However, they might also be given some kind of interpretation as n_{ei} increases with sample size and decreases as p increases. Further n_{e2} increases as ρ^2 increases, an effect that might be given the

interpretation as increased population size for estimating $\boldsymbol{\beta}_2$ by borrowing strength from the observations with missing data.

The constants, c_{i2}, $i = 1, 2$, are functions of the population coefficient of determination, R_i^2, and the intercept term, β_{0i}. The relationship between these constants and both input arguments are positive, though not linear.

$$\begin{aligned} c_{i1} =& (n_i p - 2)/[n_i (n_i - p - 1)], \quad c_{i2} = (n_1 + 1) R_i^2/\left[n_1 \left(1 - R_i^2\right)\right] + \beta_{0i}^2, \;\; i = 1, 2 \\ n_{e1} =& 1/c_{11}, \quad n_{e2} = \sigma_{22}/\left[\left(\sigma_{22} - \sigma_{12}^2/\sigma_{11}\right) c_{21} + \left(\sigma_{12}^2/\sigma_{11}\right) c_{11}\right]; \end{aligned} \tag{4}$$

Theorem 2: Prediction error over all calibration samples

The expected squared prediction errors over all calibration samples and new observations using the estimator $\tilde{\boldsymbol{\beta}}_i$, for $i = 1, 2$ under the assumptions specified above, are:

$$E\left(R_{\tilde{y}_{Ni}}\right) = \sigma_{ii}\left[1 + k_i^2 n_{ei}^{-1} + (1 - k_i)^2 c_{i2}\right]; \tag{5}$$

The proof is deferred to the Appendix.

Corollary 1:

The values for k_i minimizing the expected squared prediction error, and the corresponding expected risk functions are:

$$k_{Oracle\ i} = c_{i2}/\left(n_{ei}^{-1} + c_{i2}\right), \quad E\left(R_{\tilde{y}_{Ni}}\right)_{Oracle} = \sigma_{ii}\left\{1 + c_{i2}/\left[n_{ei}\left(n_{ei}^{-1} + c_{i2}\right)\right]\right\};$$

Corollary 2:

The values $k_{lim\ i} = \left(n_{ei}^{-1} - c_{i2}\right)/\left(n_{ei}^{-1} + c_{i2}\right)$, has the property that for $k_{lim\ i} < k_i < 1$, then ~~$E\left(R_{\tilde{y}_{Ni}}\right) < E\left(R_{\hat{y}_{Ni}}\right)$~~ $E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right) < E_{\boldsymbol{x}}\left(R_{\hat{y}_{Ni}}\right)$, where $\hat{y}_{Ni}$ denotes the prediction based on $\hat{\boldsymbol{\beta}}_{vi}$.

Corollary 3:

The expected ratios of the expected squared prediction errors of the estimator $\tilde{\boldsymbol{\beta}}_i$ and the two competitors $\hat{\boldsymbol{\beta}}_{vi}$ and $\hat{\boldsymbol{\beta}}_{2i}$, note that for $i = 1$ those are equal, the ratios of the expected prediction risks, are:

$$\begin{aligned} E\left(R_{\tilde{y}_{Ni}}\right)/E\left(R_{\hat{y}_{Ni}}\right) =& 1 - n_{ei}^{-2}/\left[\left(n_{ei}^{-1} + 1\right)\left(n_{ei}^{-1} + c_{i2}\right)\right], \\ E\left(R_{\tilde{y}_{Ni}}\right)/E\left(R_{\hat{y}_{2Ni}}\right) =& 1 - \left(c_{i1} n_{ei}^{-1} + c_{i2} n_{ei}^{-1} - c_{i1} c_{i2}\right)/\left[(c_{i1} + 1)\left(n_{ei}^{-1} + c_{i2}\right)\right]; \end{aligned}$$

Corollary 4:

The expected ratio of the expected squared prediction errors based on the estimators $\hat{\boldsymbol{\beta}}_{vi}$ *and* $\hat{\boldsymbol{\beta}}_{2i}$ *equals* 1 *for* $i = 1$ *and has a specially nice expression for* $i = 2$*, where* ~~$\rho = \sigma_{12}\left(\sigma_{11}\sigma_{22}\right)^{-1/2}$~~ $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ *is the correlation between the error terms.*

$$E\left(R_{\hat{y}_{Ni}}\right)/E\left(R_{\hat{y}_{2Ni}}\right) = 1 - \rho^2(c_{21} - c_{11})/(1 + c_{21});$$

Even though, in general, we assume $\boldsymbol{\Sigma}$ to be known in this paper the prediction risk for the estimator

$$\hat{\boldsymbol{\beta}}_{Q2} = \hat{\boldsymbol{\beta}}_{22} + q_{12}/q_{11}\left(\hat{\boldsymbol{\beta}}_{11} - \hat{\boldsymbol{\beta}}_{21}\right), \quad \boldsymbol{Q} = \left[\boldsymbol{Y}_2 - \boldsymbol{X}_2\left(\hat{\boldsymbol{\beta}}_{21}\ \ \hat{\boldsymbol{\beta}}_{22}\right)\right]^T\left[\boldsymbol{Y}_2 - \boldsymbol{X}_2\left(\hat{\boldsymbol{\beta}}_{21}\ \ \hat{\boldsymbol{\beta}}_{22}\right)\right];$$

might be analysed analytically. The fraction q_{12}/q_{11} is an unbiased estimator for σ_{12}/σ_{11} as can be derived using Giri (2003).

Lemma 1:

The expected risk function for prediction error, i.e. ~~$E\left(R_{\hat{y}_{QN2}}\right)$~~ $E_{\boldsymbol{x}}\left(R_{\hat{y}_{QN2}}\right)$*, using* $\hat{\boldsymbol{\beta}}_{Q2}$ *as estimator for* $\boldsymbol{\beta}_2$ *is:*

$$E\left(R_{\hat{y}_{QN2}}\right) = \sigma_{22}\left\{(1 + c_{21}) - (c_{21} - c_{11})\left[\rho^2 - \left(1 - \rho^2\right)/(n_2 - p - 2)\right]\right\}; \tag{6}$$

The proof is deferred to the Appendix.

Corollary 5:

The expected ratio of the expected squared prediction errors based on the estimators $\hat{\boldsymbol{\beta}}_{Q2}$ *and* $\hat{\boldsymbol{\beta}}_{22}$ *is:*

$$E\left(R_{\hat{y}_{QN2}}\right)/E\left(R_{\hat{y}_{2Ni}}\right) = 1 - \left[(c_{21} - c_{11})(1 + c_{21})\right]\left[\rho^2 - \left(1 - \rho^2\right)/(n_2 - p - 2)\right];$$

Do note that if $\rho^2 < 1/(n_2 - p - 1)$, then the expected prediction risk using the standard OLS estimator, $\hat{\boldsymbol{\beta}}_{22}$, is smaller than using the estimator $\hat{\boldsymbol{\beta}}_{Q2}$.

5 Results

As shown in ~~Fig.~~ Figure 1, the gain of using $\hat{\boldsymbol{\beta}}_{v2}$ over $\hat{\boldsymbol{\beta}}_{22}$ may be substantial, especially for combinations when ρ^2 and $(c_{21} - c_{11})/(1 + c_{21})$, are both high. The latter expression is basically reflecting the relative difference between n_1 and n_2. Figure 2 shows that further improvements might be achieved by substituting $\hat{\boldsymbol{\beta}}_{vi}$ with $\tilde{\boldsymbol{\beta}}_i$, for $i = 1, 2$ in situations when c_{i2}, basically reflecting the size of R_i^2, is small. The effect diminishes when n_{ei} increases, i.e. when the effective sample size is large.

The results of (3) and (6) were validated via simulations using the software "R" (R Core Team, 2014) and an extension of the package "simrel" (Sæbø, 2015; Sæbø et al., 2015), capable of producing a bivariate response variable. Figure 3 shows different simulation tests. As this study is not a simulation study, we contented ourselves with simulating results for a bundle of combinations for n_1 and n_2, varied ρ and plotted the results onto the theoretical risks like shown in Figure 3 for visual validation.

6 Discussion

Our major finding in this study is to show that for linear regression with bivariate response including missing data, there exists an unbiased ~~weighted least square~~ GLS estimator, $\hat{\boldsymbol{\beta}}_{v2}$, which reduces the expected prediction error compared with the standard OLS estimator, when the covariance structure of error terms is assumed to be known. The prediction precision might be further improved by shrinking the ~~weighted~~ generalized least squares estimator by the principles outlined by James and Stein (1961).

The natural next step, which is the topic of our companion paper (Gangsei et al., 2016b), is to test out estimators that do not assume known covariance structure ($\boldsymbol{\Sigma}$) and known coefficients of determination (R_i^2). In (Gangsei et al., 2016b) an empirical Bayes estimator is evaluated. A main topic is to modify the estimator $\hat{\boldsymbol{\beta}}_{Q2}$ by adding shrinkage to the term q_{12}/q_{11}. ~~The obvious choice corresponding to $\tilde{\boldsymbol{\beta}}$ is some kind of empirical Bayes estimator. Their~~ The connection ~~to~~ between empirical Bayes estimators and the James–Stein estimator

is well documented by a series of papers by Efron and Morris (1971, 1972b,a, 1973, 1975, 1976). Other candidates would be restricted maximum likelihood estimators, corresponding to $\hat{\boldsymbol{\beta}}_v$, and possibly an extension of the (C)PLS estimator (Indahl et al., 2009), capable of utilizing information from observations with missing data.

The generalisation of assuming predictors to be multivariate normal distributed might be severely biased in a lot of practical situations, especially when experiments are designed. However, for many, perhaps the majority, of practical situations, the assumption might be justified at least after some normalizing transformation of variables. The validity of the theoretical results when the assumption of normally distributed predictors is violated, has been tested for the OLS estimators by simulating results using randomly distributed, not normal distributed predictors. The effect of non-normality was found to be negligible. This seems intuitively correct, as the principles of the central limit theorem should also be applicable for the current situation.

A Appendix

A.1 Proof of Theorem 1

Since $\boldsymbol{x}_N$ is centred we may write $R_{\tilde{y}_{Ni}} = \sigma_{ii} + R_{\boldsymbol{x}_N^T\tilde{\boldsymbol{\beta}}_i}$. Further since $\hat{\boldsymbol{\beta}}_{vi}$ is normally distributed, so are $\tilde{\boldsymbol{\beta}}_i$ and $\boldsymbol{x}_n^T\left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)$ for $i = 1, 2$. We have

$$E\left[\boldsymbol{x}_N^T\left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)\right] = (1-k_i)\,\boldsymbol{x}_n^T\boldsymbol{\beta}_i, \quad var\left[\boldsymbol{x}_n^T\left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)\right] = k_i^2\boldsymbol{x}_N^T\left(\boldsymbol{X}_{(+)}^T\boldsymbol{\Sigma}_{(+)}^{-1}\boldsymbol{X}_{(+)}\right)^{-1}\boldsymbol{x}_N;$$

It might be shown that:

$$\boldsymbol{x}_N^T\left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{x}_N = 1/n_1 + \boldsymbol{z}_N^T\left(\boldsymbol{Z}_1^T\boldsymbol{Z}_1\right)^{-1}\boldsymbol{z}_N, \quad \boldsymbol{x}_N^T\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1}\boldsymbol{x}_N = 1/n_2 + \boldsymbol{z}_N^T\left(\boldsymbol{Z}_{2c}^T\boldsymbol{Z}_{2c}\right)^{-1}\boldsymbol{z}_N;$$

, where both $\boldsymbol{Z}_{2c}$ and $\boldsymbol{z}_N$ are centred with respect to the n_2 first rows, and $\boldsymbol{X}_2$ is centred with respect to all n_1 rows. Then, since

$$R_{\boldsymbol{x}_N^T\tilde{\boldsymbol{\beta}}_i} = var\left[\boldsymbol{x}_N^T\left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)\right] + \left[E\boldsymbol{x}_N^T\left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)\right]^2$$

we get (3).

A.2 Proof of Theorem 2

By applying the rules for double expectations, we may write $E\left(R_{\tilde{y}_{Ni}}\right) = E_{\boldsymbol{X}_1}\left[E_{\boldsymbol{x}_N|\boldsymbol{X}_1}\left(R_{\tilde{y}_{Ni}}\right)\right]$. Due to the assumption of independent normal distribution of the rows of $\boldsymbol{Z}_1$, we have that $\boldsymbol{Z}_1^T\boldsymbol{Z}_1$ and $\boldsymbol{Z}_{2c}^T\boldsymbol{Z}_{2c}$ are two Wishard distributed variables with scale matrix $\boldsymbol{\Gamma}^{-1}$ and $n_1 - 1$ and $n_2 - 1$ degrees of freedom, respectively (Mardia et al., 1979). Thus, their inverse matrices are inverse Wishard distributed with the same parameters. Due to the centring of $\boldsymbol{z}_N$, we have that $\boldsymbol{z}_N$ is multivariate normally distributed with zero mean and covariance matrix $[(n_i+1)/n_i]\,\boldsymbol{\Gamma}$ when centred using all rows ($i = 1$) or just the n_2 first rows ($i = 2$).

By using rules for quadratic terms (Petersen and Pedersen, 2012), and the rules for expectation of the trace, we find

$$E\left[\boldsymbol{z}_N^T\left(\boldsymbol{Z}_1^T\boldsymbol{Z}_1\right)^{-1}\boldsymbol{z}_N\right] = [(n_1+1)/n_1]\,[(p-1)/(n_1-p-1)] = c_{11} - 1/n_1$$
$$E\left[\boldsymbol{z}_N^T\left(\boldsymbol{Z}_{2c}^T\boldsymbol{Z}_{2c}\right)^{-1}\boldsymbol{z}_N\right] = [(n_2+1)/n_2]\,[(p-1)/(n_2-p-1)] = c_{21} - 1/n_2$$

Further, $E\left(\boldsymbol{x}_N^T\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T\boldsymbol{x}_N\right) = E\left(\beta_{0i}^2\right) + E\left(\boldsymbol{z}_N^T\boldsymbol{\beta}_{zi}\boldsymbol{\beta}_{zi}^T\boldsymbol{z}_N\right)$ since $E\left(\boldsymbol{z}_N\right) = \boldsymbol{0}_{(p-1)}$ for $i = 1, 2$. Finally, the term $E\left(\boldsymbol{z}_N^T\boldsymbol{\beta}_{zi}\boldsymbol{\beta}_{zi}^T\boldsymbol{z}_N\right) = \sigma_{ii}R_i^2/(1-R_i^2)$ is given by definition.

Then (5) is obtained by substituting the elements in the expressions in (3) by the general terms shown above. The corollaries are given without further proof as they are easily derived mostly by minimizing functions with respect to k_1 and k_2.

A.3 Proof of Lemma 1

Conditional on known $\boldsymbol{Q}$ it might be shown by matrix algebra and the means of the multivariate normal distribution that the distribution of $\hat{\boldsymbol{\beta}}_{Q2}$ is:

$$\hat{\boldsymbol{\beta}}_{Q2} \sim N_p\left\{\boldsymbol{\beta}_2,\ \sigma_{22}\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1} - \left(2\sigma_{12}q_{12}/q_{11} - \sigma_{11}q_{12}^2/q_{11}^2\right)\left[\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\right]\right\}$$

Due to the properties of the Normal–gamma distribution, we know that (Giri, 2003):

$$E\left(q_{12}/q_{11}\right) = \sigma_{12}/\sigma_{11}, \quad E\left(q_{12}^2/q_{11}^2\right) = \sigma_{12}^2/\sigma_{11}^2 + |\boldsymbol{\Sigma}|/\left(\sigma_{11}^2\left(n_2-p-2\right)\right);$$

Then, since $var_Q\left[E\left(\hat{\boldsymbol{\beta}}_{Q2}\right)\right] = \mathbf{0}_p^T\mathbf{0}_p$ we find $var\left(\hat{\boldsymbol{\beta}}_{Q2}\right) = E_Q\left[var\left(\hat{\boldsymbol{\beta}}_{Q2}\right)\right]$, leading to (6).

References

Bock, M. (1975). Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Ann. Statist.*, 3:209–218.

Box, G. and Tiao, G. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley series in behavioral science: quantitative methods. Addison-Wesley Pub. Co.

Brown, P. J. and Zidek, J. V. (1980). Adaptive Multivariate Ridge Regression. *Ann. Statist.*, 8:64–74.

Efron, B. and Morris, C. (1971). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part I: The Bayes Case. *J. Amer. Statist. Assoc.*, 66(336):807–815.

Efron, B. and Morris, C. (1972a). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, 59(2):335–347.

Efron, B. and Morris, C. (1972b). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part II: The Empirical Bayes Case. *J. Amer. Statist. Assoc.*, 67(337):130–139.

Efron, B. and Morris, C. (1973). Stein's Estimation Rule and its Competitors-An Empirical Bayes Approach. *J. Amer. Statist. Assoc.*, 68(341):117–130.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *J. Amer. Statist. Assoc.*, 70(350):311–319.

Efron, B. and Morris, C. (1976). Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Ann. Statist.*, 4:11–21.

Gangsei, L., Kongsro, J., Olsen, E., Røe, M., Alvseike, O., and Sæbø, S. (2016a). Prediction precision for lean meat percentage in Norwegian pig carcasses using Hennessy grading probe 7: Evaluation of methods emphasized at exploiting additional information from computed tomography. *Acta Agriculturae Scandinavica, Section A-Animal Science*, pages 1–8.

Gangsei, L. E., Almøy, T., and Sæbø, S. (2016b). Linear regression with bivariate response variable containing missing data. An empirical Bayes strategy to increase prediction precision. Submitted manuscript to Communications in Statistics – Simulation and Computation.

Giri, N. C. (2003). *Multivariate Statistical Analysis: Revised and Expanded*, volume 171. CRC Press.

Helland, I. S. and Almøy, T. (1994). Comparison of Prediction Methods when Only a Few Components are Relevant. *J. Amer. Statist. Assoc.*, 89(426):583–591.

Indahl, U. G., Liland, K. H., and Næs, T. (2009). Canonical partial least squares - a unified PLS approach to classification and regression problems. *Journal of Chemometrics*, 23(9):495–504.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic press.

Matsuda, T. and Komaki, F. (2015). Singular value shrinkage priors for Bayesian prediction. *Biometrica*, 102(4):843–854.

Minka, T. (2000). Bayesian linear regression. Technical report, Microsoft Research Cambridge.

Petersen, K. B. and Pedersen, M. S. (2012). The matrix cookbook. Technical report, University of Waterloo.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sæbø, S. (2015). *simrel: Linear Model Data Simulation and Design of Computer Experiments*. R package version 1.1-0.

Sæbø, S., Almøy, T., and Helland, I. S. (2015). simrel-A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemometrics and Intelligent Laboratory Systems*.

Upton, G. and Cook, I. (2014). *A Dictionary of Statistics 3e*. Oxford university press.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science & Business Media.

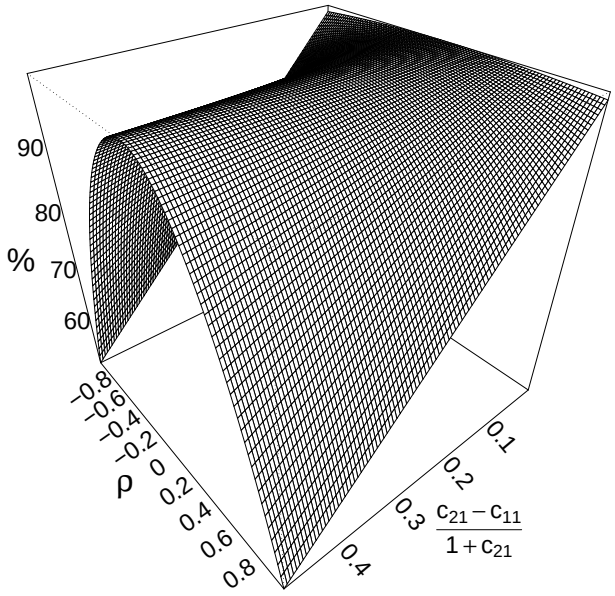

Figure 1: The expected relative size (in %) for the expected squared prediction errors using the estimator $\hat{\boldsymbol{\beta}}_{v2}$ compared with $\hat{\boldsymbol{\beta}}_{22}$ as a function of ρ and the fraction $c_{21} - c_{11}/1 + c_{21}$. As this fraction decreases, it basically means that the relative difference between the sample-sizes n_1 and n_2 also decreases.

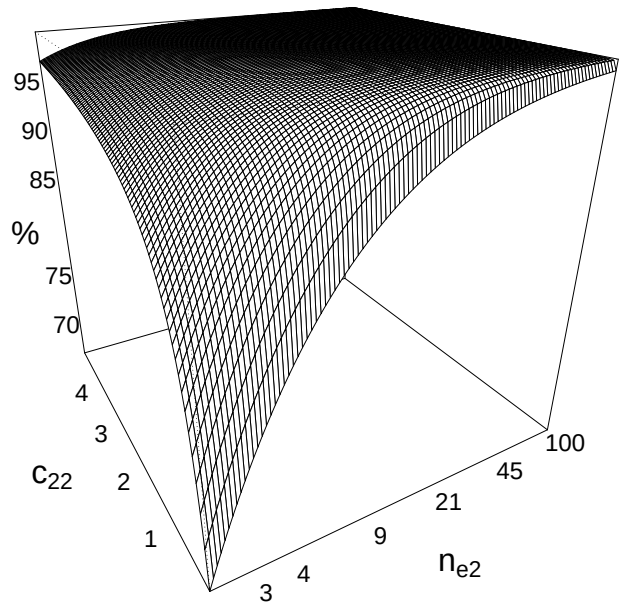

Figure 2: The expected relative size (in %) for the expected squared prediction errors using the estimator $\tilde{\boldsymbol{\beta}}_2$ compared with $\hat{\boldsymbol{\beta}}_{v2}$ as a function of n_{e2} and c_{22}. c_{22} is basically a function of R_2^2 and increases when R_2^2 increases.

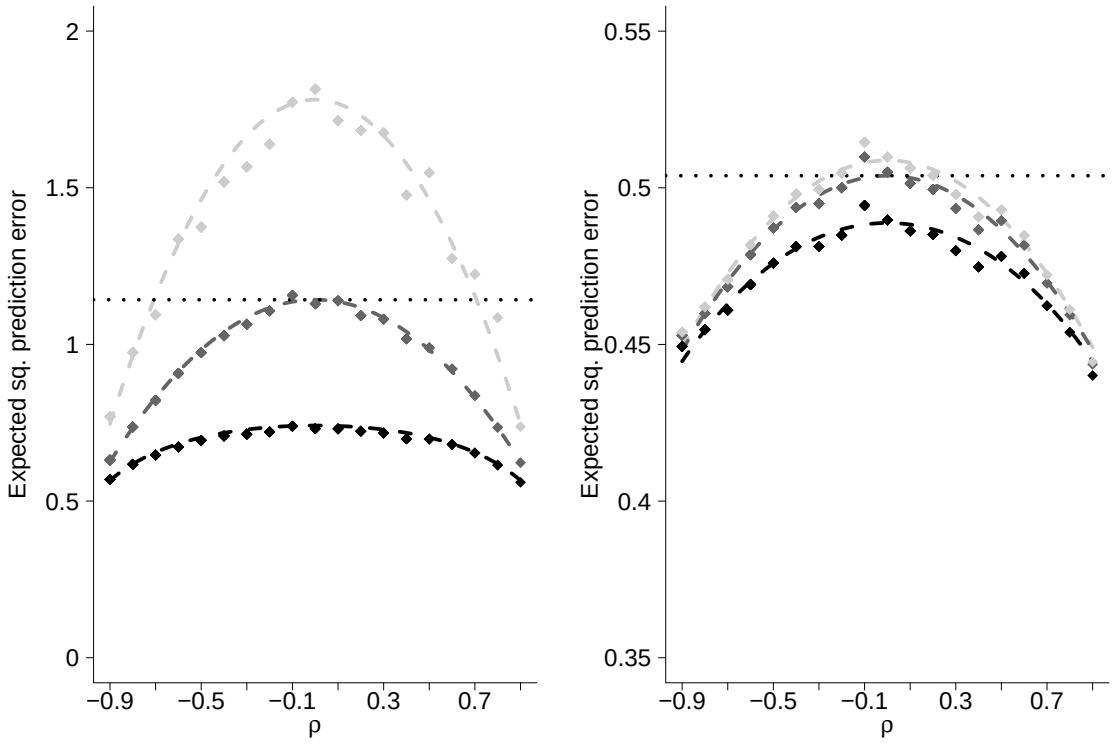

Figure 3: Expected squared prediction error using the predictors $\tilde{\boldsymbol{\beta}}_2$ (black dashed lines) $\hat{\boldsymbol{\beta}}_{v2}$ (dark gray dashed lines), $\hat{\boldsymbol{\beta}}_{Q2}$ (light gray dashed lines) and $\hat{\boldsymbol{\beta}}_{22}$ (black dotted line) as functions of ρ. n_1 equal to 20 in left panel and 50 in right panel, and n_2 equal to 7 in left panel and 20 in right panel. The simulated means are shown by diamonds in the colors corresponding to the theoretical lines. The simulation means are based on 5×10^3 independent calibration sets for each of $\rho = -0.9, -0.8, \ldots, 0.9$, and the estimates from each calibration set is used to predict 10^3 new independent observations. For all panels and simulations $p = 4$, $R_1^2 = 0.4$, $R_2^2 = 0.6$ and $\beta_{01} = \beta_{02} = 0$.