Augmenting data with published results in Bayesian linear regression

C. de Leeuw, I. Klugkist

Utrecht University, The Netherlands

Abstract

In most research, linear regression analyses are performed without taking into account published results of similar previous studies. Although the prior density in Bayesian linear regression could accommodate such prior knowledge, formal models for doing so are absent from the literature. The goal of this paper is therefore to develop a Bayesian model in which a linear regression analysis on current data is augmented with the reported regression coefficients of previous studies. Two versions of this model are presented. The first version incorporates previous studies through the prior density and is applicable when the current and all previous studies are exchangeable. The second version models all studies in a hierarchical structure and is applicable when studies are not exchangeable. Both versions of the model are assessed using simulation studies. Performance for each in estimating the regression coefficients is consistently superior to using current data alone, and is close to that of an equivalent model that uses the data from previous studies rather than reported regression coefficients. Overall the results show that augmenting data with results from previous studies is viable and yields significant improvements in the parameter estimation.

Keywords: Bayesian analysis, linear regression, informative priors, hierarchical modeling

1. Introduction

Although science itself is cumulative in nature, this is often not reflected in statistical analysis. Even when numerous studies have already been published on a given topic, new data relating to that topic is usually still analysed in isolation. This is unfortunate, as results from previous studies constitute a source of relevant information. By ignoring this information both the stability and the precision of the parameter estimates are lower than they could have been, and the conclusions that can be drawn are less certain and more likely to be affected by sampling variation. Due to the frequent combination of multiple testing with relatively low statistical power this becomes especially problematic in the framework of null hypothesis significance testing. As Maxwell (2004) demonstrates, in such a situation even exact replications of studies are likely to yield conflicting conclusions. To improve both parameter estimates and consistency in the literature it would thus be beneficial if new studies could take into account the results of relevant studies that preceded them.

In the social and behavioral sciences a commonly used model is multiple linear regression, which almost invariably includes separate significance tests for all regression coefficients. Combined with the often rather modest sample sizes, this makes it likely for different studies to find different sets of statistically significant predictors. Moreover, confidence intervals tend to be too wide to draw firm conclusions. The best way to address both these issues would be to use larger samples, but this is rarely a realistic option. Augmenting the data of a new study with previous results from similar studies might therefore be an attractive alternative, since this effectively increases the sample size as well. The goal of this paper is therefore to develop and test a formal method for augmenting data in linear regression analyses in this manner.

An obvious candidate for incorporating previous results in a new study is the prior density in a Bayesian analysis. As every textbook on Bayesian statistics is quick to point out, the posterior of one analysis can readily be used as the prior for another. It is therefore surprising that a search through the statistical literature turns up exceedingly few papers directly relevant to this issue. The topic of constructing informative

priors based on published results appears to have received little attention in Bayesian statistical research. One rare exception is Yu et al. (2009), who used previously published studies to create informative priors for the coefficients of a logistic regression. Their efforts do not constitute a comprehensive or generalizable approach however. A systematic method is not presented, and informative priors were specified for only some of the regression coefficients and separately for each of those. In general, although it is likely that ad hoc ways of specifying informative priors using literature have been used in specific applications, more formal statistical approaches to this issue are simply absent from the literature. This is especially troublesome when considering regression models, given that regression coefficients are affected by what other predictors were used. As such, the results from studies using different sets of predictors are not directly comparable, and combining those results is therefore a far from straightforward matter.

Two approaches closely related to the present issue have received more attention. The first is combining published results through meta-analysis. Though meta-analysis does not directly incorporate data, it is conceivable that the reported regression coefficients of various linear regressions could be combined using meta-analysis to obtain a prior density for a new data set. However, most research on meta-analysis deals with the univariate case (but see Nam, Mengersen & Garthwaite (2003) and Hripsime & Raudenbush (1996) for examples of multivariate meta-analysis), where an application to the present issue would call for simultaneously combining multiple regression parameters per study. Although separate meta-analyses could be done for each regression coefficient, would mean ignoring the relation between the regression coefficients (see also Riley (2009)). Finally, as already noted combining results from regression analyses with different sets of predictors presents some unique difficulties, which have not yet been addressed in the meta-analytical literature.

The second related approach is the construction of informative priors using historical data, in which the data of previous studies are used to construct the prior (see for example Ibrahim & Chen (2000)). Using the data itself rather than the reported summary statistics is clearly preferable, but obtaining that data may be too time-consuming, if it is possible at all. If a method using only the summary statistics can be shown to yield an adequate approximation of using the data itself, this would therefore provide an acceptable solution when data from previous studies are not (realistically) obtainable.

What will be presented in this paper is a Bayesian model for multiple linear regression for augmenting the data of a current study with published results of earlier studies. Two versions of the model are developed to suit two different situations. The first version incorporates previous studies through the prior, and is applicable when studies are exchangeable. The second version has a two-level hierarchical structure with current and previous data on the lower level. This version of the model is appropriate for situations where the studies are not exchangeable, and also allows study-level variables to be incorporated in the model. The two versions of the model are hereafter referred to as the replication model and hierarchical replication model respectively. For both versions it is assumed that all previous studies use the same set of predictors as the analysis on the current data. This assumption is further addressed in the discussion. Simulation studies will be performed to assess the performance of both versions of the model. The replication model is described and evaluated in Section 2, the hierarchical replication model is described and evaluated in Section 3. A discussion of the results and suggested extensions to the model is given in Section 4.

2. Replication model

2.1. Model development

In this section a Bayesian model will be developed for incorporating results from earlier studies in a new linear regression analysis through the prior density of the regression coefficients. This model is for use in situations where studies are exchangeable, that is when there is no *relevant* information other than the data or reported regression coefficients by which the studies can be distinguished. As noted, it is assumed that current data and previous studies all have the same set of predictors.

In its general form, a Bayesian analysis consists of three parts. The first part is the probability density function for the data given the unknown model parameters; taken as a function of those model parameters, this is referred to as the likelihood function. The second part is the prior density function of the model

parameters, which quantifies what is assumed to be known about the model parameters prior to observing the data. Together these yield the posterior density of the model parameters, which reflects the updated knowledge about the model parameters after having observed the data. For the replication model these three elements are described in the following three subsections, with the linear regression model for the data in Section 2.1.1; the prior densities for the regression model parameters in Section 2.1.2; and the joint posterior density as well as the Gibbs sampler used to estimate it in Section 2.1.3. The simulation study used to evaluate the performance of the replication model is described in Section 2.2.

2.1.1. The linear regression model

Let the current data set consist of an outcome variable y and a predictor data matrix X containing v predictor variables x_1 through x_v. An additional dummy variable x_0 of all 1's is included in X to obtain the intercept. For $i = 1, \cdots, n$ observations, a schematic of the data is shown in Table 1.

Table 1: Schematic of current data (X and y)

x_0	x_1	x_2	$\cdots$	x_v	y
1	$x_{1,1}$	$x_{2,1}$	$\cdots$	$x_{v,1}$	y_1
1	$x_{1,2}$	$x_{2,2}$	$\cdots$	$x_{v,2}$	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	$x_{1,n}$	$x_{2,n}$	$\cdots$	$x_{v,n}$	y_n

The model used for this data is the standard multiple linear regression model. For the ith observation,

$$y_i = \alpha_0 + \alpha_1 x_{1,i} + \cdots + \alpha_v x_{v,i} + \varepsilon_i = X_i^T \alpha + \varepsilon_i \tag{1}$$

with

$$\varepsilon_i \sim \mathrm{N}(0, \sigma_e^2), \tag{2}$$

where ε_i is the error term and the parameters $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_v)^T$ and σ_e^2 are the vector of regression coefficients and the residual variance respectively. In this model X is considered fixed, and under the usual assumption that the error terms are independent from each other it therefore follows that

$$y|_{X,\alpha,\sigma_e^2} \sim \mathrm{MVN}(X\alpha, \sigma_e^2 I_n), \tag{3}$$

with I_n the n by n identity matrix. The full Bayesian regression model is given as

$$p(\alpha, \sigma_e^2|X, y) \propto p(y|X, \alpha, \sigma_e^2) p(\alpha, \sigma_e^2). \tag{4}$$

With the probability density function for the data $p(y|X, \alpha, \sigma_e^2)$ already defined in (3), this leaves the prior density $p(\alpha, \sigma_e^2)$ to be specified.

2.1.2. The prior density

The reported results from previous studies will be incorporated in the prior for α. It is assumed that for each of $j = 1, \cdots, m$ previous studies, the regression coefficients $b_{0,j}$ through $b_{v,j}$ as well as the associated standard errors $s_{0,j}$ through $s_{v,j}$ have been reported. These are combined into the m by $v+1$ data matrices B and S respectively. Schematics for both are given in Table 2. It is further assumed that the model used in the previous studies is the linear regression model described in Section 2.1.1 (Equations (1) and (2)).

Following common convention α and σ_e^2 are assumed independent in the prior, thus:

$$p(\alpha, \sigma_e^2) = p(\alpha) p(\sigma_e^2). \tag{5}$$

For the residual variance it is assumed that no relevant prior information is available. Therefore, the prior for σ_e^2 will be specified as $p(\sigma_e^2) \propto (\sigma_e^2)^{-1}$, a common choice of non-informative prior for the variance parameter

Table 2: Schematic of B and S

b_0	b_1	b_2	$\cdots$	b_v
$b_{0,1}$	$b_{1,1}$	$b_{2,1}$	$\cdots$	$b_{v,1}$
$b_{0,2}$	$b_{1,2}$	$b_{2,2}$	$\cdots$	$b_{v,2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$b_{0,m}$	$b_{1,m}$	$b_{2,m}$	$\cdots$	$b_{v,m}$

s_0	s_1	s_2	$\cdots$	s_v
$s_{0,1}$	$s_{1,1}$	$s_{2,1}$	$\cdots$	$s_{v,1}$
$s_{0,2}$	$s_{1,2}$	$s_{2,2}$	$\cdots$	$s_{v,2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$s_{0,m}$	$s_{1,m}$	$s_{2,m}$	$\cdots$	$s_{v,m}$

of a normal distribution (for more details on common non-informative prior densities, see for example Lynch (2007) or Gelman, Carlin, Stern and Rubin (2004)).

To incorporate B and S in the prior of α, it is defined as the posterior of a second Bayesian model as follows:

$$p(\alpha) = p(\alpha|B, S) \propto p(B|S, \alpha)p^*(\alpha), \tag{6}$$

where

$$p(B|S, \alpha) = \prod_{j=1}^{m} p(B_j|S_j, \alpha). \tag{7}$$

The probability function $p^*(\alpha)$ is the *initial* prior for α, which reflects what is known about α before B and S are observed. Nothing relevant is assumed to be known initially, therefore it is specified as a non-informative improper uniform density over the whole real line, $p^*(\alpha) \propto 1$ (see for example Lynch (2007) or Gelman et al. (2004) for more details).

As (6) suggests the data matrix S is fixed in this model, leaving only the probability density for B to be specified. As the sampling distribution of regression coefficients in a Least Squares regression analysis is asymptotically normal the multivariate normal distribution with mean vector α is used for the regression coefficients B_j for each study j.

With S, only the estimated standard errors of the reported regression coefficients are known; the correlations between coefficients are rarely reported. These correlations could be included in the model as unknown parameters to be estimated, but the number of previous studies will typically be too small to obtain reliable estimates. Since furthermore the correlations are not themselves of interest, they are all restricted to 0. A simulation study (not reported in this paper) was performed to determine the effect of this restriction relative to using more accurate values, but no notable differences in model performance were found. With this restriction the covariance matrix for the distribution of B_j becomes

$$\mathbb{S}_j = \begin{pmatrix} s_{0,j}^2 & 0 & 0 & \cdots & 0 \\ 0 & s_{1,j}^2 & 0 & \cdots & 0 \\ 0 & 0 & s_{2,j}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & s_{v,j}^2 \end{pmatrix},$$

and therefore

$$B_j|_{S_j,\alpha} \sim \text{MVN}(\alpha, \mathbb{S}_j), \tag{8}$$

which completes the specification of the model in (6).

2.1.3. Posterior density and Gibbs sampler

With the above specification for the prior the following joint posterior density is obtained:

$$p(\alpha, \sigma_e^2|X, y, B, S) \propto p(y|X, \alpha, \sigma_e^2) \prod_{j=1}^{m} p(B_j|S_j, \alpha) p^*(\alpha) p(\sigma_e^2)$$

$$\propto (\sigma_e^2)^{-(\frac{n}{2}+1)} \exp\left\{ -\frac{1}{2\sigma_e^2}(y-X\alpha)^T(y-X\alpha) - \frac{1}{2}\sum_{j=1}^{m}(B_j-\alpha)^T\mathbb{S}_j^{-1}(B_j-\alpha) \right\}. \quad (9)$$

A Gibbs sampler is a convenient choice for estimating this joint posterior. The Gibbs sampler is an iterative algorithm in which random draws are made in each iteration from the conditional posteriors of the parameters given the most recently sampled values of the other parameters (see Lynch (2007) for more details). For the replication model the Gibbs sampler consists of the following steps:

Step 0 Initialize iteration counter c at c = 0. Assign a starting value $\alpha^{(0)}$.

Step 1 Set $c = c + 1$.

Step 2 Set $\sigma_e^{2(c)}$ to a random draw from $p(\sigma_e^2|X, y, \alpha^{(c-1)})$.

Step 3 Set $\alpha^{(c)}$ to a random draw from $p(\alpha|X, y, B, S, \sigma_e^{2(c)})$.

Step 4 Return to step 1.

Integrating (9), the conditional posteriors used for the Gibbs sampler are

$$\sigma_e^2|_{X,y,\alpha} \sim \text{inv-gamma}\left(\frac{n}{2}, \frac{1}{2}(y-X\alpha)^T(y-X\alpha)\right) \quad (10)$$

for step 2, and

$$\alpha|_{X,y,B,S,\sigma_e^2} \sim \text{MVN}\left(\mu_\alpha, \Sigma_\alpha\right) \quad (11)$$

for step 3, where

$$\mu_\alpha = \left(\frac{X^TX}{\sigma_e^2} + \sum_{j=1}^{m}(\mathbb{S}_j)^{-1}\right)^{-1}\left(\frac{X^Ty}{\sigma_e^2} + \sum_{j=1}^{m}(\mathbb{S}_j)^{-1}B_j\right)$$

$$\Sigma_\alpha = \left(\frac{X^TX}{\sigma_e^2} + \sum_{j=1}^{m}(\mathbb{S}_j)^{-1}\right)^{-1}.$$

2.2. Model evaluation

2.2.1. Method

A simulation study was performed to assess model performance. For the simulations, data were generated from a multivariate normal distribution containing six predictor variables x_1 to x_6 and an outcome variable y. For each simulation one sample was generated to serve as the current data. To simulate the previous studies, samples were generated from the same population. An Ordinary Least Squares regression using the model described in Section 2.1.1 was then applied to each of these samples to obtain the estimated regression coefficients B and the corresponding standard errors S. Five simulation parameters were manipulated one at a time to determine their impact on model performance, for a total of 11 simulations. A total of 100 data sets was generated for each simulation.

The specification of the population and sampling is as follows. The means and standard deviations for the x's were 0 and 1 respectively; y had a mean of 5 and a standard deviation of 2. The correlations between

$x_1, \cdots, x_6$ and y were $0.1, \cdots, 0.6$ respectively. The default values for the simulation parameters that were varied were 0.3 for the bivariate correlations between x's, 50 for the current data sample size and previous study sample size, 3 for the number of previous studies, and 4 for the number of predictors. Alternative values were 0.1 and 0.5 for the bivariate correlations, 25 and 100 for the sample sizes, 1 and 5 for the number of previous studies, and 2 and 6 for the number of predictors. For simulations with 4 predictors, x_1, x_2, x_4 and x_5 were used. For the simulation with 2 predictors x_1 and x_4 were used. Only one simulation parameter was varied at a time, with default values for the four other parameters.

In addition to the replication model, two more models were applied to each data set. The first is the current data model, the Bayesian linear regression model as described in Section 2.1.1 applied to just the current data, with an unbounded uniform prior for α, so $p(\alpha) \propto 1$. The comparison with this model serves to illustrate the potential gain from including previous studies in the prior. The second additional model is the full data model, which is the same as the current data model but applied to the combination of current data and the data of the previous studies. This model represents the ideal scenario where all data are available, instead of just the published summaries for the previous studies.

2.2.2. Results

For all simulations, two measures of model performance were considered. For the first, the posterior means of α were stored for each of the three models. The difference between the posterior means of α for the replication model and current data model on the one hand and those of the full data model on the other hand was computed. Taking the full data model as the best-case scenario, this difference in posterior means gives a measure of how close the performance of the other two models is to this best-case scenario. The mean and 95% interval over all data sets of the difference in posterior means was calculated for each simulation. Results are shown for the simulation with 6 predictors and default values for the other simulation parameters in Figure 1.

As can be seen in the figure the means of the difference are close to zero everywhere, which indicates that there are no biases. The widths of the 95% interval are reasonably small for the replication model, meaning the posterior means of the replication model tend to stay close to those of the full data model. For the current data model however the variability is much greater, with the difference in posterior means reaching 0.4 in some cases. Given the fact that the regression coefficients for the predictors are about 0.5 on average, these constitute very large divergences. As the figure also shows there are no differences between predictors, from which it follows that the strength of correlations between the x's and y have no effect. The same results were also found in the other simulations, with no appreciable effect of the simulation parameters that were varied. It can therefore be concluded that in this respect the performance of the replication model is both clearly superior to the standard approach using just the current data, and a reasonable approximation of

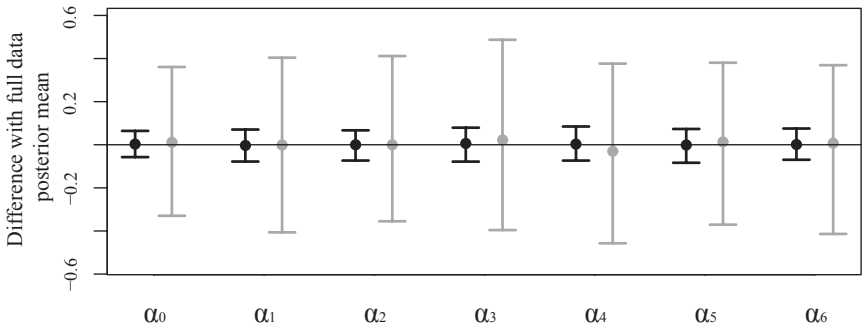

Figure 1: Interval plot by predictor of the difference between the posterior mean of the full data model and the posterior mean of the replication model (black lines) and the posterior mean of the current data model (gray lines). The error bars show the 95% interval of the difference, the dots are the mean difference. The shown results are for the simulation with 6 predictors and default values for all other simulation parameters.

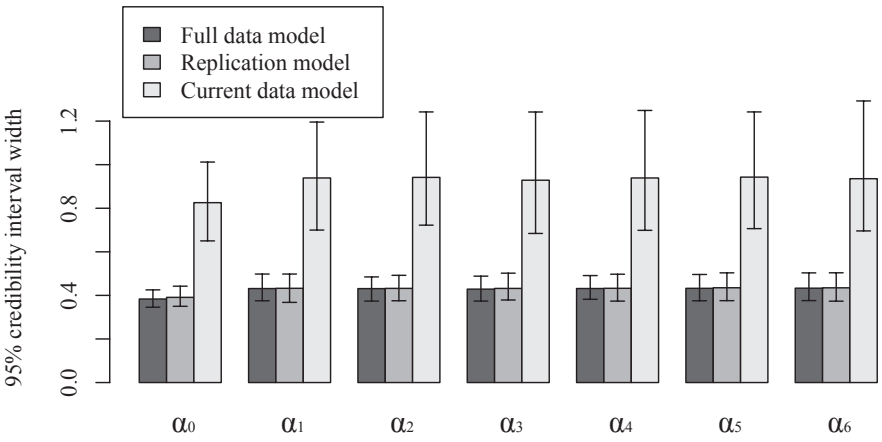

Figure 2: Bar chart of the mean width of the 95% credibility interval (CI) for each predictor, for each of the three models. The error bars show the 95% interval of the CI width. The shown results are for the simulation with 6 predictors and default values for all other simulation parameters.

the performance in the best-case scenario with all data available.

The second measure of model performance that was used was the width of the 95% credibility interval (CI) for α. Where the first measure of model reflected accuracy of the posterior estimates, the CI widths reflect the certainty of those estimates. The mean and 95% interval over all data sets of the CI widths were calculated for each simulation, and they are shown for the simulation with 6 predictors and default values for the other simulation parameters in Figure 2. From the figure it is apparent that there is no real difference in CI widths between the full data model and replication model. Both the mean and the variability of the CI widths are essentially the same. Performance of the current data model is quite poor in comparison, with mean CI widths more than twice as large as those of the other two models and significantly greater variability.

The differences between predictors are again negligible, but here there are differences across simulations. Performance for the full data model and replication model remained roughly identical in all conditions, but the performance of the current data model relative to the other two models did change. For each data set the difference in CI widths for α between the current data model and the replication model were computed, and the mean and 95% interval of this difference over all data sets was then calculated for each simulation. These are shown in Figure 3 for α_4, grouped by the simulation parameters that were varied. All simulation parameters except the number of predictors appear to systematically affect the CI widths. For the first three simulation parameters this is easily explained: as the current data size decreases, or the number of previous studies or their sample size increases, the amount of information contained in the previous studies relative to the current data becomes larger. Consequently it should be no surprise that the precision of the replication model, which uses that information, increases relative to the current data model, which does not. The effect of the increase in correlations between predictor variables is harder to explain, but it might be that the loss of information in the sample due to those larger correlations more severely affects the current data model.

Even though the relative performance of the current data model is better in some of the simulations, it never reaches the level of the replication model. With CI widths for the replication model on average about half the size of the current data model, its performance in this respect is again consistently superior to that of the current data model. Generally, what the simulations demonstrate is that the replication model offers a significant improvement with regard to parameter estimation over using the current data only. Moreover, the differences between the replication model and the best-case scenario with all data available are minor. This suggests that in those circumstances were the replication model is applicable the gain in using all data is likely not worth the effort needed to obtain them.

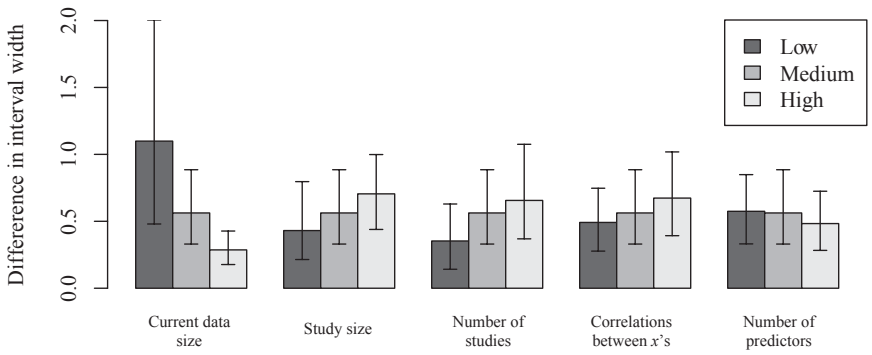

Figure 3: Bar chart of the mean difference in width of the 95% credibility intervals (CI) for α_4 between the current data model and the replication model, grouped by simulation parameter. Each set of bars correspond to the three values for that simulation parameter given in Section 2.2.1, with the values of all other simulations at their default value. The error bars show the 95% interval of the CI width.

3. Hierarchical replication model

3.1. Model development

3.1.1. Likelihood and prior density

For the replication model developed in Section 2 all studies were assumed to be exchangeable, which is reflected in the fact that the current data and the results from the previous studies were all used to estimate one common parameter vector α. In this section a hierarchical version of the replication model is developed for situations where studies are not exchangeable. This is the case in situations where there is relevant information besides the data and reported regression coefficients by which studies can be distinguished.

In the hierarchical replication model each study j estimates its own set of regression coefficients $\beta_j = (\beta_{0j}, \beta_{1j}, \cdots, \beta_{vj})^T$, with $j = 0$ for the current data and with v again the number of predictor variables. Together these form the distribution for the data

$$p(y, B|X, S, \beta_0, \beta_1, \cdots, \beta_m, \sigma_e^2) = p(y|X, \beta_0, \sigma_e^2) \prod_{j=1}^{m} p(B_j|S_j, \beta_j), \tag{12}$$

with for the current data

$$y|_{X,\beta_0,\sigma_e^2} \sim \text{MVN}(X\beta_0, \sigma_e^2 I_n) \tag{13}$$

and for the previous studies, $j = 1, \cdots, m$,

$$B_j|_{S_j,\beta_j} \sim \text{MVN}(\beta_j, \mathbb{S}_j). \tag{14}$$

Furthermore, the studies are considered to have been drawn from a distribution with unknown parameters

$$p(\beta_0, \beta_1, \cdots, \beta_m|Z, \gamma, \Lambda) = \prod_{j=0}^{m} p(\beta_j|Z_j, \gamma, \Lambda), \tag{15}$$

where for $j = 0, \cdots, m$

$$\beta_j|_{Z_j,\gamma,\Lambda} \sim \text{MVN}(\gamma Z_j, \Lambda). \tag{16}$$

Here, Z_j is a vector of w study-level predictor variables for study j, with an additional dummy $z_{j0} = 1$ to obtain the study-level intercept. These study-level predictors are used to account for differences between studies. Correspondingly, γ is the $v + 1$ by $w + 1$ matrix of study-level regression parameters. Each row k of γ is associated with the regression coefficients β_{kj}, and each column l is associated with the study-level

predictor z_l. If no study-level predictor variables are used, the product γZ_j reduces to a vector of means for β_j that is the same for all j.

Under the usual assumption of independent priors, the prior density for the remaining parameters is

$$p(\sigma_e^2, \gamma, \Lambda) = p(\sigma_e^2)p(\gamma)p(\Lambda). \tag{17}$$

For σ_e^2 the same prior is used as before, $p(\sigma_e^2) \propto \sigma_e^{-2}$. Each element of γ is given a uniform prior over the whole real line, so $p(\gamma) \propto 1$. To assign a prior to the covariance matrix Λ it is first decomposed into a matrix Ω and a vector ξ, $\Lambda = \text{Diag}(\xi)\Omega\text{Diag}(\xi)$. Here Ω and ξ roughly correspond to the correlation matrix and standard deviations respectively, although Ω is not constrained to be a correlation matrix. This strategy allows priors to be specified for the correlation matrix and the standard deviations separately. The priors defined in this manner are generally less constraining on Λ than the standard inverse Wishart prior for the covariance matrix Λ would be. See Barnard, McCulloch & Meng (2000) for more details on this separation approach. The priors for the components of Λ are

$$\Omega \sim \text{Inv-Wishart}_{v+2}(I_{v+1}) \tag{18}$$

and

$$\xi_k \sim \text{LogN}(0, 1.5^2) \tag{19}$$

for $k = 0, \cdots, v$.

3.1.2. Posterior density and hybrid Gibbs sampler

Combining all elements described in Section 3.1.1, the joint posterior density for the model becomes

$$p(\beta_0, \beta_1, \cdots, \beta_m, \sigma_e^2, \gamma, \Lambda | X, y, B, S, Z) \propto p(y,B|X,S,\beta_0,\beta_1, \cdots, \beta_m,\sigma_e^2)p(\beta_0,\beta_1, \cdots, \beta_m|Z,\gamma,\Lambda)p(\sigma_e^2,\gamma,\Lambda). \tag{20}$$

As a Gibbs sampler is again used to estimate this joint posterior density, the conditional posterior densities for all parameters need to be derived. Substituting β_0 for α, the conditional posterior for σ_e^2 is the same as in (10). For β_0 it is

$$\beta_0|_{X,y,Z_0,\sigma_e^2,\gamma,\Lambda} \sim \text{MVN}\left(\mu_{\beta_0}, \Sigma_{\beta_0}\right), \tag{21}$$

with

$$\mu_{\beta_0} = \left(\frac{X^T X}{\sigma_e^2} + \Lambda^{-1}\right)^{-1} \left(\frac{X^T y}{\sigma_e^2} + \Lambda^{-1}\gamma Z_0\right)$$
$$\Sigma_{\beta_0} = \left(\frac{X^T X}{\sigma_e^2} + \Lambda^{-1}\right)^{-1}.$$

For all β_j, $j \neq 0$,

$$\beta_j|_{B_j,S_j,Z_j,\gamma,\Lambda} \sim \text{MVN}\left(\mu_{\beta_j}, \Sigma_{\beta_j}\right), \tag{22}$$

with

$$\mu_{\beta_j} = \left(\mathbb{S}_j^{-1} + \Lambda^{-1}\right)^{-1} \left(\mathbb{S}_j^{-1} B_j + \Lambda^{-1}\gamma Z_j\right)$$
$$\Sigma_{\beta_j} = \left(\mathbb{S}_j^{-1} + \Lambda^{-1}\right)^{-1}.$$

For each column $l = 0, \cdots, w$ of γ the conditional posterior is

$$\gamma_l|_{Z,\beta_0,\beta_1,\cdots,\beta_m,\gamma_{(-l)},\Lambda} \sim \text{MVN}\left(\frac{\sum_{j=0}^m Z_{jl}(\beta_j - \gamma_{(-l)}Z_{j(-l)})}{\sum_{j=0}^m Z_{jl}^2}, \frac{\Lambda}{\sum_{j=0}^m Z_{jl}^2}\right), \tag{23}$$

and for Ω, the conditional posterior is

$$\Omega|Z, \beta_0, \beta_1, \cdots, \beta_m, \gamma, \xi \sim \text{Inv-Wishart}_{m+v+3}\left(I_{v+1} + (\text{Diag}(\xi))^{-1}\,\Delta\,(\text{Diag}(\xi))^{-1}\right) \tag{24}$$

with

$$\Delta = \sum_{j=0}^{m} (\beta_j - \gamma Z_j)(\beta_j - \gamma Z_j)^T.$$

Finally, for each ξ_k, $k = 0, \cdots, v$, the conditional posterior is

$$p(\xi_k|Z,\beta_0,\beta_1,\cdots,\beta_m,\gamma,\Omega,\xi_{(-k)}) \propto \xi_k^{-(m+2)} \exp\left\{ -\frac{\Delta_{kk}^2 (\Omega^{-1})_{kk}}{2\xi_k^2} - \frac{\sum_{i\neq k}^{v} (\Omega^{-1})_{ik}\Delta_{ik}/\xi_i}{\xi_k} - \frac{(\log \xi_k)^2}{4.5} \right\}. \quad (25)$$

As this does not reduce to a known parametric form it cannot be sampled from directly. Therefore, a Metropolis-Hastings algorithm is used to separately sample each ξ_k conditional on the other parameters, implemented as a step in the Gibbs sampler. This Metropolis-Hastings step has the following substeps, with c the current iteration of the Gibbs sampler:

Step 1 Set ξ_k^* to a random draw from the proposal density $p_{\text{pr}}(.)$.

Step 2 Set u to a random uniform draw between 0 and 1.

Step 3 Compute the ratio $R = \frac{p(\xi_k^*|Z,\beta_0,\beta_1,\cdots,\beta_m,\gamma,\Omega,\xi_{(-k)})}{p(\xi_k^{(c-1)}|Z,\beta_0,\beta_1,\cdots,\beta_m,\gamma,\Omega,\xi_{(-k)})}$.

Step 4 If $R > u$, set $\xi_k^{(c)} = \xi_k^*$. Otherwise, set $\xi_k^{(c)} = \xi_k^{(c-1)}$.

Step 5 Return to step 1.

For proposal density $p_{\text{pr}}(.)$ the normal distribution with mean at $\xi_k^{(c-1)}$ and variance τ_k^2 is used. The value of τ_k^2 is tuned during the burn-in of the Gibbs sampler such that the acceptance rate for the Metropolis-Hastings step gets close to 0.44, the optimal acceptance rate for univariate MH sampling (see Gelman, Roberts & Gilks (1996)).

3.2. Model evaluation

3.2.1. Method

A simulation study was done to illustrate the performance of the hierarchical replication model. Data was generated with four predictor variables x_1 to x_4 and an outcome variable y, but unlike the simulations in Section 2.2 it was assumed that further relevant information about the current and previous studies was known by which they could be distinguished. As an example, the gender ratio for the sample of each study is used here, represented by the sample proportion of males denoted z_1. The variable *gender* itself is unobserved in the sample, only the ratio for the whole sample is known. Consequently, if there are gender differences in the effects of $x_1, \cdots, x_4$ on y, results of studies would differ according to the gender ratio of the sample used. Studies with different gender ratios would therefore have to be considered not to be exchangeable, as their samples are in effect drawn from different populations.

For the simulation study, the number of previous studies was set to three. Sample sizes for both the current data and previous studies was set to 50. For the sake of simplicity, the values for z_1 were not randomly sampled but were specified in advance, with a value of 0.4 for the current data and 0.2, 0.6 and 0.9 for the previous studies. Two multivariate normal populations for the predictor and outcome variables were defined, one for males and one for females. Data for each of the studies were generated by drawing samples from these separate populations in the appropriate proportion, as given by z_1. For the previous studies, Ordinary Least Squares linear regression was again used to obtain the regression coefficients and their standard errors.

The populations were defined as follows. In both, the means and standard deviations of the x's were 0 and 1 respectively and the standard deviation of y was 2. The mean of y was 5 for the male population and

Table 3: Correlations for the male and female populations, with differences highlighted in bold font

	Male					Female				
Predictor	x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4	y
x_1	1	**0.1**	**0.1**	0	0.1	1	**0.4**	**0.4**	0	0.1
x_2	**0.1**	1	**0.1**	0	0.3	**0.4**	1	**0.4**	0	0.3
x_3	**0.1**	**0.1**	1	0	0.5	**0.4**	**0.4**	1	0	0.5
x_4	0	0	0	1	**0.1**	0	0	0	1	**0.4**
y	0.1	0.3	0.5	**0.1**	1	0.1	0.3	0.5	**0.4**	1

3 for the female population. The correlations for the two populations are given in Table 3, with differences between the populations highlighted in bold font. The simulation was run for a total of 100 data sets.

Two comparison models were run alongside the hierarchical replication model. The first was the current data model, the same as in Section 2.2. The second was the hierarchical full data model. This again represents the ideal situation where data from previous studies are available. This model has the same specification as the hierarchical replication model, except that the data of the previous studies are used instead of just B and S, using the same data distribution (13) as the current data. In the hierarchical full data model all studies estimated the same, shared residual variance parameter σ_e^2.

3.2.2. Results

For this simulation the same two measures of model performance were used as in the simulations in Section 2.2. The first one was the difference in posterior means of β_0 of the hierarchical replication model and current data model on the one hand, and those of the hierarchical full data model on the other hand. Computing the mean and 95% interval of these differences over all data sets gives a measure of the accuracy of the hierarchical replication model and current data model. The second measure of model performance was the width of the 95% credibility intervals of β_0. For these the mean and 95% interval of these widths over all data sets is computed for each of the three models. This gives a measure of the certainty of the posterior estimates of the models.

The results for the first measure of model performance, the difference of posterior means with those of the hierarchical full data model, are shown in Figure 4. The results are very similar to those for the nonhierarchical replication model. Again the posterior means for both models are similar on average to those of the full data model, but as before the variability was much greater for the current data model than for the hierarchical replication model. No differences were found between the predictors. Thus, as with the nonhierarchical version of the replication model, the hierarchical replication model shows a clear advantage

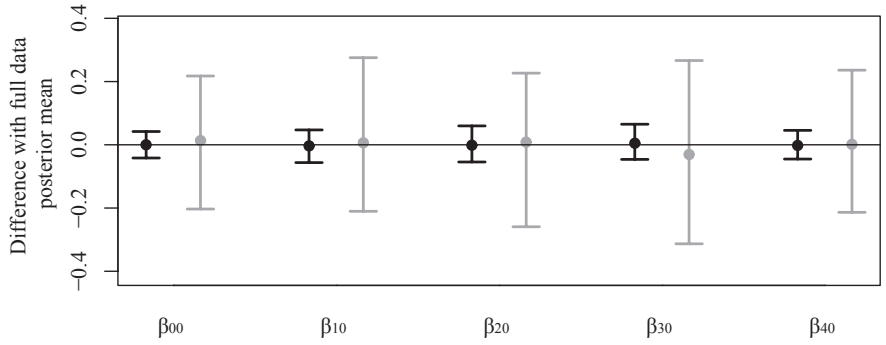

Figure 4: Interval plot for β_0 by predictor of the difference between the posterior mean of the full data model and the posterior mean of the replication model (black lines) and the posterior mean of the current data model (gray lines). The error bars show the 95% interval of the difference, the dots are the mean difference.

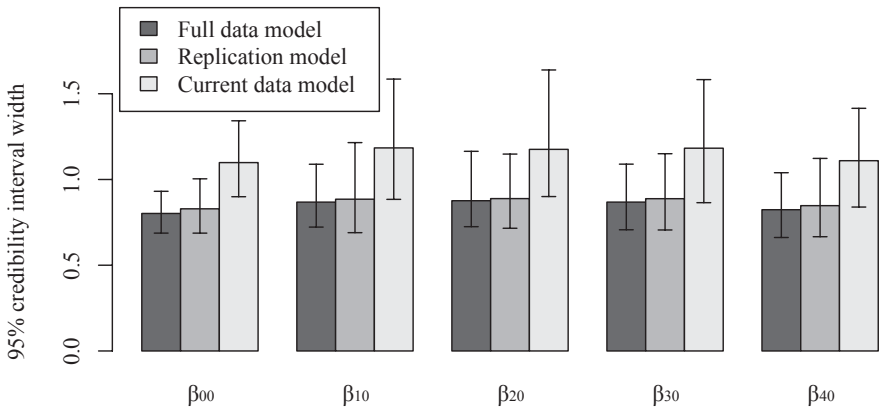

Figure 5: Bar chart for β_0 of the mean width of the 95% credibility interval (CI) for each predictor, for each of the three models. The error bars show the 95% interval of the CI width.

over using the current data only in this respect. In addition, it also reasonably approximates the performance of the hierarchical full data model.

Results for the second measure of model performance, the widths of the 95% credibility intervals (CI) for β_0, are shown in Figure 5. Compared to the simulations in Section 2.2 the improvement due to using the previous studies shows the same pattern as before, but the effect is not as large. This is in part a result of the fact that in this case the studies all come from somewhat different populations, as explained in Section 3.2.1. Consequently the results of each previous study reflects the particular population it is drawn from, and these results are therefore less informative about the population of the current study. As such the certainty of the estimates of β_0 does not improve as much when including the previous results as it did for the estimates of α in Section 2.2. What does remain the same is that the hierarchical replication and full data models show little difference for both the mean and the variability of the CI width, and both do still perform better than the current data model for all predictors.

Whatever conclusions follow from the estimates of β_0, these are essentially restricted to the specific population from which the current study is drawn, with a particular gender ratio. Clearly, if gender has no effect on the regression coefficients the conclusions would be more general in their scope. Yet in this case, when using just the current data it cannot be determined whether gender indeed has no effect (and in fact, in this simulation it does have an effect), since gender is not itself included as a predictor variable. It follows that when using the current data model, it is not really possible to determine what the scope of the conclusions actually is. The researcher would either have to draw very limited conclusions, or must assume that gender has no effect. In this regard the hierarchical replication model offers an additional advantage. Besides superior estimates of β_0, the hierarchical replication model also provides information about the relation between the different studies. By looking at the estimates of γ, the research can gain an insight into how the regression coefficients of the different studies are related and what role the gender ratio plays in any differences between those studies.

To assess the performance of the hierarchical replication model with respect to estimating γ, the posterior means of γ_0 and $\gamma_0 + \gamma_1$ were compared to the posterior means of the hierarchical full data model as well as the true population means. The reason $\gamma_0 + \gamma_1$ rather than γ_1 was used is that $\gamma_0 + \gamma_1$ estimates the mean regression coefficients for a completely male population, in the same way that γ_0 does so for a completely female population. Using $\gamma_0 + \gamma_1$ therefore makes it easier to compare the estimates to the true population means. In Figure 6 are shown for $\gamma_0 + \gamma_1$ the mean and 95% interval over all data sets of the difference between the posterior means with those true population means. As can be seen in the figure neither of the models shows any bias, and the variability of the estimates was the same for both. The same applies to γ_0. No difference between the models was found for the widths of the 95% credibility intervals (not shown in a plot here) of both γ_0 and $\gamma_0 + \gamma_1$ either. Thus, in addition to offering improved estimates of β_0

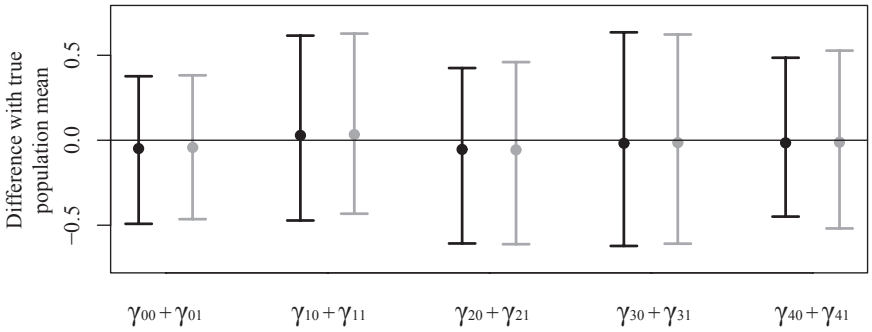

Figure 6: Interval plot for $\gamma_0 + \gamma_1$ by predictor of the difference between the true population means and both the posterior mean of the full data model (black lines) and the posterior mean of the replication model (gray lines). The error bars show the 95% interval of the difference, the dots are the mean difference. For all k, l, γ_{kl} is the regression parameter of z_l for the coefficients of the kth predictor variable x_k.

compared to the current data model, the hierarchical replication model also provides reliable information on the potential differences between studies as well as to what extent those differences can be explained by study-level variables.

4. Discussion

The goal of this paper was to develop a model for augmenting a multiple linear regression analysis on new data with published results from earlier studies. Two questions that were addressed with respect to the model were: how much would be gained in comparison to using just the new data; and how large would the difference be with the ideal situations in which data from previous studies were available. For both questions the results of the simulation studies were encouraging. Performance of the two versions of the replication model was consistently superior to using current data alone, with less variability in posterior means and smaller credibility intervals. Moreover, both versions of the model showed performance equivalent to models that used all the data of the previous studies rather than just reported results.

It can thus be concluded that the model developed in this paper offers the possibility of obtaining significantly better parameter estimates in a linear regression setting, without needing to expend a prohibitive amount of time and energy trying to obtain data from the previous studies. Moreover, the hierarchical version of the model offers the advantage of being able to address questions about differences between studies. In particular, it gives explicit information on the extent to which the estimates based on the current data are representative for a more general population. This in contrast to the usual approach of linear regression on just the current data, where this question is usually implicitly ignored and findings are potentially generalized to a greater extent than would be appropriate.

One important issue that still needs to be addressed is the restriction that all studies have the same set of predictors. It was imposed in order to facilitate model development, and the next logical step in the development would be to remove it in order to make the model applicable to a wider range of situations. Removing this restriction does pose a challenging problem, as the reported regression coefficients in previous studies are affected by which other predictors were included in the regression. To be used in the present model the previous studies need to have used the same set of predictors as are of interest in the current study. When they do not this would require that, in addition to filling in missing regression coefficients, the reported regression coefficients are adjusted to the values the would have had, had the needed set of predictors been used.

There are other issues that remain to be addressed as well, but regardless of any limitations of the present model the conclusion is clear. Incorporating the results of previous studies in a linear regression on new data does yield significantly better parameter estimates, and it provides a more than adequate approximation of

using the actual data of those previous studies. Expanding the model to make it more generally applicable in practice is thus certainly warranted.

References

Barnard, J., McCulloch, R., & Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281 - 1311.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis.* Chapman & Hall/CRC.

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 5, proceedings of the fifth valencia international meeting* (p. 599 - 607). Oxford: Clarendon Press.

Hripsime, A. K., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*(3), 227-235.

Ibrahim, J., & Chen, M. (2000). Power prior distributions for regression models. *Statistical S*, *15*(1), 46 - 60.

Lynch, S. M. (2007). *Introduction to applied bayesian statistics and estimation for social scientists.* Springer.

Maxwell, S. (2004). The persistence of underpowered studies in psychological research: causes consequences, and remedies. *Psychological Methods*, *9*(2), 147 - 163.

Nam, I., Mengersen, K., & Garthwaite, P. (2003). Multivariate meta-analysis. *Statistics in Medicine*, *22*, 2309 - 2333.

Riley, R. D. (2009). Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society*, *172*, 789-811.

Yu, X., Xun, P., Hu, Z., Liu, P., Shen, H., & Chen, F. (2009). Combining previously published studies with current data in bayesian logistic regression model: an example for identifying susceptibility genes related to lung cancer in humans. *Journal of Toxicology and Environmental Health*, *72*, 683 - 689.