Open Journal of Statistics, 2015, 5, 680-701
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.57069

Scientific Research Publishing

On the Covariance of Regression Coefficients

Pantelis G. Bagos*, Maria Adam

Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
Email: *pbagos@compgen.org

Received 2 October 2015; accepted 14 December 2015; published 17 December 2015

Abstract

In many applications, such as in multivariate meta-analysis or in the construction of multivariate models from summary statistics, the covariance of regression coefficients needs to be calculated without having access to individual patients' data. In this work, we derive an alternative analytic expression for the covariance matrix of the regression coefficients in a multiple linear regression model. In contrast to the well-known expressions which make use of the cross-product matrix and hence require access to individual data, we express the covariance matrix of the regression coefficients directly in terms of covariance matrix of the explanatory variables. In particular, we show that the covariance matrix of the regression coefficients can be calculated using the matrix of the partial correlation coefficients of the explanatory variables, which in turn can be calculated easily from the correlation matrix of the explanatory variables. This is very important since the covariance matrix of the explanatory variables can be easily obtained or imputed using data from the literature, without requiring access to individual data. Two important applications of the method are discussed, namely the multivariate meta-analysis of regression coefficients and the so-called synthesis analysis, and the aim of which is to combine in a single predictive model, information from different variables. The estimator proposed in this work can increase the usefulness of these methods providing better results, as seen by application in a publicly available dataset. Source code is provided in the Appendix and in http://www.compgen.org/tools/regression.

Keywords

Meta-Analysis, Linear Regression, Covariance Matrix, Regression Coefficients, Synthesis Analysis

1. Introduction

The linear regression model is one of the oldest and most commonly used models in the statistical literature and it is widely used in a variety of disciplines ranging from medicine and genetics to econometrics, marketing, social sciences and psychology. Moreover, the relations of the linear regression model to other commonly used

*Corresponding author.

How to cite this paper: Bagos, P.G. and Adam, M. (2015) On the Covariance of Regression Coefficients. *Open Journal of Statistics*, **5**, 680-701. http://dx.doi.org/10.4236/ojs.2015.57069

methods such as the *t*-test, the Analysis of Variance (ANOVA) and the Analysis of Covariance (ANCOVA) [1] [2], as well as the role played by the multivariate normal distribution in multivariate statistics, place the linear model in the centre of interest in many fields of statistics.

In several applications, expressions for estimates of various parameters of the multiple regression models in terms of the summary statistics are needed. This is more evident in the general area of research synthesis methods, in which a researcher seeks to combine multiple sources of evidence across studies. For instance, in meta-analysis of regression coefficients [3], which is a special case of multivariate meta-analysis [4] [5], one is interested in the covariance matrix of the coefficients obtained in various studies, in order to perform a multivariate meta-analysis that takes properly into account the correlations among the estimates. The synthesis of regression coefficients has received increased attention in recent years [3]. This growing interest is probably related to the increasing complexity of the models investigated in primary research, and this seems to be the case for both biological [6] [7] as well as social sciences [8]-[11]. However, as Becker and Wu point out in their work: "*the covariance matrix among the slopes in primary studies is rarely reported* (*though matrices of correlations among predictors are sometimes reported*)" [3]. A well-known result from linear regression theory suggests that the covariance matrix of the coefficients depends on the cross-product matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, where $\boldsymbol{X}$ is the design matrix of the independent variables. Thus, in such a case, one needs to have access to individual data, something which is difficult and time-consuming.

Another example is the case of the so-called "synthesis analysis", the aim of which is to combine in a single predictive model information from different variables. Synthesis analysis differs from traditional meta-analysis, since we are not synthesizing similar outcomes across different studies, but instead, we are trying to construct a multivariate model from pairwise associations, or to update a previously created model using external information (*i.e.* for an additional variable). For example, let's consider the case of a multiple linear regression model that relates the de-pendent variable, *y*, with *p* independent variables $x_1, x_2, \cdots, x_p$. The aim of the method is to build the multivariate model that relates all predictors, however, not the individual data, but rather the information arising from the pairwise relationships among the variables. Samsa and coworkers were the first to provide details of such method. They used the univariate linear regressions of each x_i against *y* and the correlation matrix that describes the linear relationships among the x_i's [12]. However, they did not provide an estimate for the covariance matrix. Later, Zhou and coworkers presented a different version of the method in which they used the univariate linear regressions of each x_i against *y* along with the simple regressions that related each pair of x_i's [13]. Their method was based on solving a linear system of equations and they also described a method for calculating the variance-covariance matrix of the estimated coefficients using the multivariate delta method, utilizing the estimated variance-covariance matrix of the individual regression models. Such methods could be very important for instance for adjusting a previously obtained estimate for a potential confounder, for adjusting the results of a new analysis using estimates from the literature [14], or for constructing and updating multivariate risk models [15]-[17].

In this work, we derive an analytic expression for the covariance matrix of the regression coefficients in a multiple linear regression model. In contrast to the well-known expressions which make use of the cross-product matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, we express the covariance matrix of the regression coefficients directly in terms of covariance matrix of the explanatory variables. This is very important since the covariance matrix of the explanatory variables can be easily obtained, or even imputed using data from the literature, without requiring access to individual data. In the following, in the Methods section we first present the details of synthesis analysis (2.1) and meta-analysis (2.2), in order to establish notation. Then, in Section (2.3) we present the classical framework of the multivariate normal model on which the problem is based and we give some results concerning some previously published estimators. Afterwards, in Section (2.4) we present the main result consisting of the analytical expression for the covariance of the regression coefficients. Finally, in Section (3) the method is applied to a real dataset, both in a meta-analysis and a synthesis analysis framework. Source code that implements the method, as well as the derivations of the main results are given in Appendix.

2. Methods

2.1. Synthesis Analysis

The aim of synthesis analysis is to combine in a single predictive model, information from different variables.

For instance, consider the case of a multiple linear regression model that relates the dependent variable, *y*, with *p* independent variables, $x_1, x_2, \cdots, x_p$. The traditional linear regression, models the expectation of *y* given $x_1, x_2, \cdots, x_p$. as a linear combination of the covariates:

$$E\left(y \mid x_1, x_2, \cdots, x_p\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{1}$$

The aim of the method is to build the model in Equation (1), in other words, to find the estimates of the parameters $\beta_0, \beta_1, \cdots, \beta_p$, using however not the individual data, but rather the information arising from the pairwise relationships among the variables. In the following, the regression coefficients are the elements of the (p + 1) × 1 matrix $\boldsymbol{\beta}^* = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix}$, where $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$. These relationships could be, from the one hand the univariate linear regressions of each x_i against *y*:

$$E\left(y \mid x_i\right) = a_{0i} + a_{1i} x_i, \ \ i = 1, 2, \cdots, p \tag{2}$$

On the other hand, we could either have the simple regressions that relate each pair of x_i's:

$$E\left(x_j \mid x_i\right) = \gamma_{0ij} + \gamma_{1ij} x_i, i \neq j, \ \ 1 \leq i, \ \ j \leq p \tag{3}$$

or, the correlation matrix that describes the linear relationships among the x_i's:

$$\boldsymbol{R}_x = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix} \tag{4}$$

In Equation (4), the Pearson's correlation coefficient between x_i and x_j are denoted by r_{ij} for $1 \leq i, j \leq p$. The first approach for synthesis analysis was presented by Samsa and coworkers [12] who used Equation (2) and Equation (3) in order to calculate the estimates of Equation (1). In particular, the authors used a previously known result that relates $\boldsymbol{\beta}$ to the matrices ***A***, ***S***, where $\boldsymbol{\beta}$ is the $p \times 1$ matrix of the regression coefficients $\beta_1, \beta_2, \cdots, \beta_p$ from the multivariate model of Equation (1), ***A*** is the $p \times 1$ matrix of the regression coefficients of Equation (2), ***S*** is the $p \times 1$ matrix of the standard deviations of the x_i covariates and $\boldsymbol{R}_x$ is given by Equation (4). If we denote ***A*** and ***S*** by:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1p} \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} s_{x_1} \\ s_{x_2} \\ \vdots \\ s_{x_p} \end{bmatrix} \tag{5}$$

then the regression coefficients can be calculated by:

$$\boldsymbol{\beta} = \frac{\boldsymbol{R}_x^{-1}\left(\boldsymbol{A} \odot \boldsymbol{S}\right)}{\boldsymbol{S}} \tag{6}$$

In Equation (6), $\odot$ stands for the element-wise multiplication (also known as the Hadamard product or dot matrix product) and similarly the division (/) is also element-wise. This method, provides estimates for the regression coefficient $\beta_1, \beta_2, \cdots, \beta_p$, and in order for the intercept, β_0, to be calculated, one would need to use the estimated $\beta_1, \beta_2, \cdots, \beta_p$ along with the mean values of the variables. Finally, we should mention that the method as described did not provide an estimate for the variance of the coefficients. Thus, construction of confidence intervals and assessment of the statistical significance of the covariates could not be carried out. In a latter work, Zhou and coworkers [13] developed a different method. First, they took expectations on both sides of Equation (1) conditioning on x_i:

$$E\left(y \mid x_i = x\right) = \beta_0 + \beta_1 E\left(x_1 \mid x_i = x\right) + \cdots + \beta_i x + \cdots + \beta_p E\left(x_p \mid x_i = x\right) \tag{7}$$

Then, by combining Equation (2), Equation (3) and Equation (7), they obtained the following result:

$$a_{0i} + a_{1i} x = \beta_0 + \left(\beta_1 \gamma_{0ij} + \cdots + \beta_k \gamma_{0ip}\right) + \left(\beta_1 \gamma_{1ij} + \cdots + \beta_i + \cdots + \beta_p \gamma_{1ip}\right) x \tag{8}$$

Using now Equation (8), they obtained a system of p equations for the p unknown parameters, which are the p elements of $\boldsymbol{\beta}$, $\beta_1, \beta_2, \cdots, \beta_p$, that can be easily solved and p equations for the intercept β_0, which however they proved that lead to a unique solution. The authors described also a method for calculating the variance-covariance matrix of the estimated coefficients using the multivariate delta method, utilizing the estimated variance-covariance matrix of the individual regression models (Equation (2) and Equation (3)).

The method is very interesting in that it does not assume normality of the covariates in order to estimate the parameters and thus it is expected to be more robust in case of non-normally distributed variables (but assumes the normality of the estimated parameters in order to use the delta method). On the other hand, the method is quite difficult to be implemented for an arbitrary number of covariates. The system of equations arising from Equation (8) should be solved explicitly and the solution will be more difficult as the number of covariates increases (the authors provided explicit solutions for p = 2 and p = 3). The major difficulty however, lies in the calculation of the covariance matrix with the delta method. The difficulty is particularly evident if we consider that the β_i's are highly non-linear functions of the α_i's and γ_i's and thus the partial derivatives require explicit calculations, which are different for different p and can be done only using software that perform symbolic calculations.

2.2. Meta-Analysis of Regression Coefficients

In the meta-analysis of regression coefficients, the problem is different. Here, we have a set $\beta_{1s}, \beta_{2s}, \cdots, \beta_{ps}$ of p regression coefficients arising from k studies ($s = 1, 2, \cdots, k$) and we want to combine them in order to obtain the overall mean $\boldsymbol{\beta}$. Thus, it is a special case of multivariate random-effects meta-analysis [4] [5]; we denote $\boldsymbol{\beta}_s = \left(\beta_{1s}, \beta_{2s}, \cdots, \beta_{ps}\right)$ and usually assume that $\boldsymbol{\beta}_s$ is distributed following a multivariate normal distribution around the true means $\boldsymbol{\beta}$, according to the marginal model:

$$\boldsymbol{\beta}_s \sim MVN\left(\boldsymbol{\beta}, \boldsymbol{\Omega} + \boldsymbol{C}_s\right) \tag{9}$$

In the above model, we denote by $\boldsymbol{C}_s$ the within-studies covariance matrix:

$$\boldsymbol{C}_s = \begin{bmatrix} s_{1s}^2 & \rho_{w_{12}} s_{2s} s_{1s} & \cdots & \rho_{w_{1p}} s_{1s} s_{ps} \\ \rho_{w_{12}} s_{1s} s_{2s} & s_{2s}^2 & \cdots & \rho_{w_{2p}} s_{2s} s_{ps} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{w_{1p}} s_{1s} s_{ps} & \rho_{w_{2p}} s_{2s} s_{ps} & \cdots & s_{ps}^2 \end{bmatrix} \tag{10}$$

and by $\boldsymbol{\Omega}$ the between-studies covariance matrix, given by:

$$\boldsymbol{\Omega} = \begin{bmatrix} \tau_0^2 & \rho_{B10} \tau_0 \tau_1 & \cdots & \rho_{Bp0} \tau_0 \tau_p \\ \rho_{B10} \tau_0 \tau_1 & \tau_1^2 & \cdots & \rho_{Bp1} \tau_1 \tau_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{Bp0} \tau_0 \tau_p & \rho_{Bp1} \tau_1 \tau_p & \cdots & \tau_p^2 \end{bmatrix} \tag{11}$$

This is the classic model of multivariate meta-analysis used in several applications [4] [5] [18]. For fitting this model, there are several alternatives, such as Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) or the multivariate method of moments (MM), all of which however require that the diagonal elements of $\mathbf{C}_s$. These are the study-specific estimates of the variance that are assumed known, whereas the off-diagonal elements correspond to the pairwise within-studies covariances, thus for $i, j = 0, 1, \cdots, p$ we have:

$$\rho_{wij} s_{is} s_{js} = \operatorname{cov}\left(\beta_{is}, \beta_{js}\right) \tag{12}$$

On the other hand, the between studies covariance matrix is estimated from the data. Of course, in model of Equation (9) we could also use $\boldsymbol{\beta}^*$ and include the intercept as well. However, this will rarely be needed in

practical applications where the interest lies in the estimation of covariate effects.

The major problem in this method, is, as Becker and Wu point out that "*in practice, the covariance matrix among the slopes in primary studies is rarely reported* (*though matrices of correlations among predictors are sometimes reported*)" [3]. Usually, ignoring or approximating the within studies covariance matrix produce reliable estimates for the fixed effects parameters but biased estimates for the variance [19] [20]. Thus, ideally one would want to include reliable estimates for the within studies covariances in order to gain the maximum from the multivariate meta-analysis. Currently, since the majority of studies do not report the covariance matrices, a literature-based (*i.e.* without having access to individual data) meta-analysis would be forced to assume zero correlations between the regression coefficients, limiting this way the efficiency of the method. An alternative, would be to use the model of Riley and coworkers, which, being no-hierarchical, maintains the individual weighting of each study in the analysis but includes only one overall correlation parameter, removing this way the need to know the within-study correlations [21]. For other effect sizes, such as the odds ratio, the relative risk and so on, recent studies have shown that under certain conditions, the correlation can be estimated using only the pairwise correlations of the variables involved [22] [23]. Thus, a similar approach can be followed here concerning the regression coefficients.

2.3. The General Method

We will begin with the multivariate normal model. This is one of the two main approaches for formulating a regression problem (the other one is the approach that assumes that the independent variables are fixed by design). Even though the two approaches are conceptually very different, it is well known that concerning the estimation of the regression parameters (coefficients and their variance), they yield exactly the same results. Consider we have $p + 1$ variables, y and $x_1, x_2, \cdots, x_p$ that are distributed according to a multivariate normal distribution. The traditional linear regression, models the expectation of y given $x_1, x_2, \cdots, x_p$ as a linear combination of the covariates $x_1, x_2, \cdots, x_p$ according to Equation (1). If we denote by $\boldsymbol{Y} = \left(y, x_1, x_2, \cdots, x_p\right)$, $\boldsymbol{\mu} = \left(\overline{y}, \overline{x}_1, \overline{x}_2, \cdots, \overline{x}_p\right)$ and by $\boldsymbol{\Sigma}$ the $(p+1)\times(p+1)$ variance-covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} s_y^2 & s_{x_1 y} & s_{x_2 y} & \cdots & s_{x_p y} \\ s_{y x_1} & s_{x_1}^2 & s_{x_2 x_1} & \cdots & s_{x_p x_1} \\ s_{y x_2} & s_{x_1 x_2} & s_{x_2}^2 & \cdots & s_{x_p x_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{y x_p} & s_{x_1 x_p} & s_{x_2 x_p} & \cdots & s_{x_p}^2 \end{bmatrix} \tag{13}$$

then we will have $\boldsymbol{Y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a well-known result from multivariate statistics allows the arbitrary partitioning of $\boldsymbol{\Sigma}$ in order to obtain:

$$\boldsymbol{Y} \sim MVN\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right) \tag{14}$$

In this case, the partial vectors are once again multivariate normal with $\boldsymbol{Y}_1 \sim MVN(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, $\boldsymbol{Y}_2 \sim MVN(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ with $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$, and $\boldsymbol{\Sigma}_{22}$ being the partial covariance matrices. Then, the conditional distribution of $\boldsymbol{Y}_1$ given $\boldsymbol{Y}_2$ (*i.e.* the regression of $\boldsymbol{Y}_2$ on $\boldsymbol{Y}_1$) is given by:

$$\boldsymbol{Y}_1 \,|\, \boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim MVN\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{y}_2 - \boldsymbol{\mu}_2\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right) \tag{15}$$

If we partition $\boldsymbol{\Sigma}$ in order to obtain Equation (1), then the partial covariance matrices would be:

$$\boldsymbol{\Sigma}_{12} = \begin{bmatrix} s_{x_1 y} & s_{x_2 y} & \cdots & s_{x_p y} \end{bmatrix}, \boldsymbol{\Sigma}_{22} = \begin{bmatrix} s_{x_1}^2 & s_{x_2 x_1} & \cdots & s_{x_p x_1} \\ s_{x_1 x_2} & s_{x_2}^2 & \cdots & s_{x_p x_2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{x_1 x_p} & s_{x_2 x_p} & \cdots & s_{x_p}^2 \end{bmatrix}, \boldsymbol{\Sigma}_{21} = \begin{bmatrix} s_{y x_1} \\ s_{y x_2} \\ \vdots \\ s_{y x_p} \end{bmatrix} \tag{16}$$

whereas, $\boldsymbol{Y}_1$ would be a univariate normal $y \sim N\left(\overline{y}, s_y^2\right)$. Then, the regression coefficients of Equation (1), with

the exception of the intercept, will be given by:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{17}$$

The intercept is given simply as a function of the p regression coefficients and the mean vectors of y and x_i's:

$$\beta_0 = \overline{y} - \sum_{i=1}^{p} \beta_i \overline{x}_i$$

The covariance matrix of the coefficients in Equation (1) including the intercept is given by:

$$\operatorname{cov}\left(\boldsymbol{\beta}^*\right) = \sigma^2 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \tag{18}$$

where $\boldsymbol{X}$ denotes the $n \times p$ design matrix of the independent variables, $\boldsymbol{X}^{\mathrm{T}}$ the transpose matrix of $\boldsymbol{X}$ and $\sigma^2 = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = s_y^2 - \boldsymbol{\Sigma}_{12}\boldsymbol{\beta}$. Notice that $\boldsymbol{\beta}^{\mathrm{T}} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$, since $\boldsymbol{\Sigma}_{22}$ is a symmetric matrix and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^{\mathrm{T}}$. An alternative estimate for $\operatorname{cov}(\boldsymbol{\beta})$ in terms of the centralised design matrix $\boldsymbol{X}_c$, is discussed in **Appendix C**.

In **Appendix A**, we show that the estimated regression coefficients by this method are identical to the ones obtained by Samsa and coworkers [12]. In other words, we show that:

$$\frac{\boldsymbol{R}_x^{-1}\left(\boldsymbol{A} \odot \boldsymbol{S}\right)}{\boldsymbol{S}} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{19}$$

It is obvious that the estimate proposed by Samsa and coworkers [12] is just a re-parameterization of a well-known result and produces identical estimates.

Another commonly used formula, can be derive the estimation of the standardised regression coefficients b_i, for each $i = 1, 2, \cdots, p$, using the correlation matrices $\boldsymbol{R}_x$ of Equation (4) and $\boldsymbol{R}_{xy} = \boldsymbol{R}_{yx}^{\mathrm{T}} = \begin{bmatrix} r_{1y} & r_{2y} & \cdots & r_{py} \end{bmatrix}$, where r_{iy} are the Pearson's correlation coefficient between x_i and y. Then, the matrix $\mathbf{b}$ of standardised regression coefficients can be obtained by:

$$\boldsymbol{b} = \boldsymbol{R}_x^{-1}\boldsymbol{R}_{yx} \tag{20}$$

The standardised regression coefficients b_i can be transformed to unstandardised regression coefficients β_i using $\beta_i = b_i\, s_y / s_{x_i}$, or in matrix form

$$\boldsymbol{\beta} = s_y \boldsymbol{V}_x^{-1}\boldsymbol{b} \tag{21}$$

where $\boldsymbol{V}_x$ denotes the $p \times p$ diagonal matrix such that $\boldsymbol{V}_x = \operatorname{diag}\left(s_{x_1}, s_{x_2}, \cdots, s_{x_p}\right)$. In **Appendix B**, we show that the coefficients obtained with Equation (20) and Equation (21), are identical to the ones obtained with the use of Equation (6) and Equation (17). That is, we show that:

$$s_y \boldsymbol{V}_x^{-1}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{yx} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{22}$$

Thus, it is clear that the three methods described above are equivalent and yield identical estimates

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = s_y \boldsymbol{V}_x^{-1}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{yx} = \frac{\boldsymbol{R}_x^{-1}\left(\boldsymbol{A} \odot \boldsymbol{S}\right)}{\boldsymbol{S}} \tag{23}$$

2.4. Variance-Covariance Matrix

If we want to obtain the variance of the estimated coefficients, we need to turn to Equation (18), which requires explicit knowledge of the $n \times p$ design matrix $\boldsymbol{X}$ and the cross-product matrix:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \begin{bmatrix} n & \sum x_1 & \cdots & \sum x_p \\ \sum x_1 & \sum x_1^2 & \cdots & \sum x_1 x_p \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_p & \sum x_p x_1 & \cdots & \sum x_p^2 \end{bmatrix} \tag{24}$$

In synthesis analysis as well as in meta-analysis, one usually does not have access to $n \times p$ individual data $\boldsymbol{X}$,

and thus, we have to find a method to estimate the variance with information obtained from the published reports. As we already discussed, Samsa and coworkers [12] did not provide an estimate for the covariance matrix, whereas Zhou and coworkers [13] provided a difficult to obtain estimate, using the delta method. However, the variance of a regression coefficient (let's say β_i) can be obtained relatively easily from summary statistics and it can be shown to be equal to:

$$\mathrm{var}\left(\beta_i\right)=\frac{\sigma^2}{\left(n-1\right)s_{x_i}^2}\frac{1}{1-R_i^2} \tag{25}$$

The formula in Equation (25) can be found in many textbooks with the proof traced back in earlier versions of Green's Econometrics Analysis [24]; another elegant proof can be found in [25]. Here, R_i^2 is the squared multiple correlation that relates x_i with the rest of the independent variables, whereas σ^2 is the total variance of the regression. The term $1/\left(1-R_i^2\right)$ is usually named "variance inflation factor". In many applications, we may conveniently assume that the total variance remains the same if we add x_i in the model, so we may write the variance of the regression coefficient in the full model as a function of the variance of the coefficient in the univariate model of Equation (2):

$$\mathrm{var}\left(\beta_i\right)\approx\mathrm{var}\left(a_{1i}\right)\frac{1}{1-R_i^2} \tag{26}$$

Clearly, in most of the situations this is an upper bound [25] [26] that leads to conservative estimates but it may be useful in many practical applications. In order to evaluate Equation (25), we need to calculate R_i^2 and σ^2. As we already said, σ^2 can be obtained from:

$$\sigma^2=\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}=s_y^2-\boldsymbol{\Sigma}_{12}\boldsymbol{\beta} \tag{27}$$

We need to remind however, that since this quantity is usually estimated, in real applications we need to adjust it (see [27] pp 405) in order to obtain the unbiased estimator:

$$\hat{\sigma}^2=\frac{n-1}{n-p-1}\left(\hat{s}_y^2-\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\beta}}\right) \tag{28}$$

On the other hand, an easy way (among others) to obtain R_i^2 is using:

$$R_i^2=1-1/\upsilon_{ii} \tag{29}$$

where υ_{ii} is the i-th diagonal element of $\boldsymbol{R}_x^{-1}$.

The main result of this work is to provide a closed-form expression for the covariance, that does not include $\boldsymbol{X}$. In **Appendix C** we show that the covariance is given by:

$$\mathrm{cov}\left(\beta_i,\beta_j\right)=-\frac{\sigma^2 r_{ij}}{\left(n-1\right)s_{x_ix_j}}\cdot\frac{r_{ij;12\cdots p}}{\sqrt{\left(1-R_i^2\right)}\sqrt{\left(1-R_j^2\right)}} \tag{30}$$

where $r_{ij;12\cdots p}$ is the ij-partial correlation of x_i with x_j-controlling for the remaining variables. For $i\neq j$, the ij-partial correlation coefficient is defined [28] as:

$$r_{ij;12\cdots p}=\left(-1\right)^{i+j+1}\frac{\det\left(\boldsymbol{R}_x\right)_{ij}}{\sqrt{\det\left(\boldsymbol{R}_x\right)_{ii}}\sqrt{\det\left(\boldsymbol{R}_x\right)_{jj}}} \tag{31}$$

with $\left(\boldsymbol{R}_x\right)_{ij}$ being the submatrix that is obtained by deleting the i-th row and j-th column of the correlation matrix $\boldsymbol{R}_x$ in Equation (4). Moreover, for $i=j$ the already known from Equation (25) variance of β_i is recovered as follows:

$$\mathrm{var}\left(\beta_i\right)=\frac{\sigma^2}{\left(n-1\right)s_{x_i}^2\left(1-R_i^2\right)}=\frac{\sigma^2\det\left(\boldsymbol{R}_x\right)_{ii}}{\left(n-1\right)s_{x_i}^2\det\boldsymbol{R}_x} \tag{32}$$

Interestingly, the correlation between the coefficients will simply be given by:

$$corr\left(\beta_i,\beta_j\right)=\frac{\operatorname{cov}\left(\beta_i,\beta_j\right)}{\sqrt{\operatorname{var}\left(\beta_i\right)\operatorname{var}\left(\beta_j\right)}}=-\frac{\dfrac{\sigma^2}{(n-1)s_{x_ix_j}}\cdot\dfrac{r_{ij}r_{ij;123\cdots p}}{\sqrt{\left(1-R_i^2\right)\left(1-R_j^2\right)}}}{\sqrt{\dfrac{\sigma^2}{(n-1)s_{x_i}^2\left(1-R_i^2\right)}}\sqrt{\dfrac{\sigma^2}{(n-1)s_{x_j}^2\left(1-R_j^2\right)}}}=-r_{ij;123\cdots p} \tag{33}$$

Thus, another useful relation can be obtained if we consider the $p\times p$ diagonal matrix $V_{\boldsymbol{\beta}}$ such that $\boldsymbol{V}_{\boldsymbol{\beta}}=\operatorname{diag}\left(\sqrt{\operatorname{var}\left(\beta_1\right)},\sqrt{\operatorname{var}\left(\beta_2\right)},\cdots,\sqrt{\operatorname{var}\left(\beta_p\right)}\right)$, then, it is obvious that

$$\operatorname{cov}\left(\boldsymbol{\beta}\right)=-\boldsymbol{V}_{\beta}\boldsymbol{P}_p\boldsymbol{V}_{\beta} \tag{34}$$

where $\boldsymbol{P}_p$ is the $p \times p$ matrix of ij-partial correlations $r_{ij;123\cdots p}$. The variance of the intercept (β_0) can be obtained by using the properties of the covariance function, some well known results from linear regression and Equation (25), Equation (30):

$$\begin{aligned}
\operatorname{var}\left(\beta_0\right)&=\operatorname{var}\left(\bar{y}-\sum_j\beta_j\bar{x}_j\mid x_1,x_2,\cdots,x_p\right)\\
&=\operatorname{var}\left(\bar{y}\mid x_1,x_2,\cdots,x_p\right)+\operatorname{var}\left(\sum_j\beta_j\bar{x}_j\mid x_1,x_2,\cdots,x_p\right)-2\operatorname{cov}\left(\bar{y},\sum_j\beta_j\bar{x}_j\mid x_1,x_2,\cdots,x_p\right)\\
&=\operatorname{var}\left(\bar{y}\mid x_1,x_2,\cdots,x_p\right)+\sum_j\bar{x}_j^2\operatorname{var}\left(\beta_j\right)+2\sum_{i,j}\operatorname{cov}\left(\beta_i,\beta_j\right)\\
&=\frac{\sigma^2}{n}+\sigma^2\sum_j\frac{\bar{x}_j^2}{(n-1)s_{x_j}^2\left(1-R_j^2\right)}-2\sigma^2\sum_{1\le j<i<p}\left\{\frac{r_{ij}}{(n-1)s_{x_jx_i}}\cdot\frac{r_{ij;12\cdots p}}{\sqrt{\left(1-R_i^2\right)}\sqrt{\left(1-R_j^2\right)}}\right\}
\end{aligned} \tag{35}$$

Similarly, we may obtain the covariance of β_0 with β_i:

$$\begin{aligned}
\operatorname{cov}\left(\beta_0,\beta_i\right)&=\operatorname{cov}\left(\bar{y}-\sum_j\beta_j\bar{x}_j,\beta\mid x_1,x_2,\cdots,x_p\right)\\
&=\operatorname{cov}\left(\bar{y},\beta_i\mid x_1,x_2,\cdots,x_p\right)-\operatorname{cov}\left(\sum_j\beta_j\bar{x}_j,\beta_i\mid x_1,x_2,\cdots,x_p\right)\\
&=-\operatorname{cov}\left(\sum_j\beta_j\bar{x}_j,\beta_i\mid x_1,x_2,\cdots,x_p\right)=-\operatorname{cov}\left(\beta_i\bar{x}_i+\sum_{j\ne i}\beta_j\bar{x}_j,\beta_i\mid x_1,x_2,\cdots,x_p\right)\\
&=-\operatorname{cov}\left(\beta_i\bar{x}_i,\beta_i\mid x_1,x_2,\cdots,x_p\right)-\operatorname{cov}\left(\sum_{j\ne i}\beta_j\bar{x}_j,\beta_i\mid x_1,x_2,\cdots,x_p\right)\\
&=-\bar{x}_i\operatorname{var}\left(\beta_i\right)-\sum_{j\ne i}\bar{x}_j\operatorname{cov}\left(\beta_j,\beta_i\right)\\
&=2\sigma^2\sum_{1\le i<j<p}\left\{\frac{\bar{x}_jr_{ij}}{(n-1)s_{x_jx_i}}\cdot\frac{r_{ij;12\cdots p}}{\sqrt{\left(1-R_i^2\right)}\sqrt{\left(1-R_j^2\right)}}\right\}-\sigma^2\frac{\bar{x}_i}{(n-1)s_{x_i}^2\left(1-R_i^2\right)}
\end{aligned} \tag{36}$$

At this point we should note that Equation (33) was also mentioned by Becker and Wu, and was attributed to [29]. However, the formula was given there only as an unsolved problem for the regression with two independent variables. Most probably, Becker and Wu (since they were aware of the formula), overlooked the fact that the partial correlation coefficient can be calculated from the pairwise correlations, using simple matrix manipulations. To the best of the authors' knowledge, Equation (30) and its derivation is novel, since it cannot be found or mentioned in any of the traditional books of linear regression or multivariate analysis [24] [27] [29]-[33].

3. Results and Discussion

As an illustrative example for both meta-analysis and synthesis analysis, we used a publicly available dataset concerning Diabetes in Pima Indians. The dataset has been created from a larger dataset and it was obtained from the UCI Machine Learning Repository [34] (http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes). The dataset has been used in the past in several applications for constructing prediction models for diabetes [35]. Here, we chose to use plasma glucose concentration at 2 hours in an oral glucose tolerance test as the dependent variable. For predictors, we used the diastolic blood pressure (mm Hg), the triceps skin fold thickness (mm), the 2-hour serum insulin (mu U/ml), the body mass index (weight in kg/(height in m)^2) and the age (years). The code provided in **Appendix D**, makes clear that the method, using only the summary statistics, produces identical estimates with the standard linear regression analysis on the original data.

Afterwards, we used the same dataset in order to create an "artificial" meta-analysis dataset. We randomly split the dataset in 10 subsets (which we treat as "studies") of approximate the same number of participants (from 55 to 86). For each dataset, we performed the same calculation and estimated the same model for predicting plasma glucose concentration. The estimates for the regression coefficients and their standard errors in each subset are listed in **Table 1**. Then, we applied the various alternative methods in a meta-analysis of these 10 "studies", in order to investigate the effect of the different within-studies covariance matrix.

Firstly, we used the actual within studies covariance matrix obtained from each dataset, which is the ideal but not easily tenable situation. Secondly, we assumed a zero within studies correlation (that is, we used only the variances of the regression coefficients). Thirdly, we applied the alternative method of Riley and coworkers [21] that does not differentiate between and within studies variation (and thus, it requires as input only the variances). And last, we applied the proposed method by assuming a realistic scenario, in which the variances of the regression coefficients are known, but the covariances are not, and thus they are imputed. For all analyses we used the mvmeta command in Stata with the REML option [36].

By observing the pooled correlation matrix between the independent variables (measured in the combined dataset of 768 individuals), which was found to be equal to:

Table 1. The estimates of the regression coefficients and their standard errors, after randomly splitting the dataset in 10 subsets (which we treat as "studies"). For each dataset, we performed the same calculation and estimated the same model for predicting plasma glucose concentration. The regression coefficients for each subset (s) correspond to diastolic blood pressure (β_1), triceps skin fold thickness (β_2), 2-hour serum insulin (β_3), body mass index (β_4) and age (β_5).

Subset (s)	β_1 (se)	β_2 (se)	β_3 (se)	β_4 (se)	β_5 (se)
1	-0.237540 (0.155266)	0.294668 (0.203763)	0.075400 (0.027186)	1.413119 (0.425623)	1.355647 (0.272464)
2	0.134445 (0.178934)	-0.433280 (0.263380)	0.092133 (0.033765)	0.599211 (0.583968)	0.135907 (0.315666)
3	0.396087 (0.244395)	-0.322338 (0.239103)	0.104948 (0.035626)	0.749045 (0.474633)	0.607717 (0.280881)
4	-0.141942 (0.171337)	-0.374355 (0.255028)	0.088818 (0.027895)	0.512126 (0.470001)	0.365304 (0.313291)
5	0.231959 (0.172719)	-0.809532 (0.294184)	0.140998 (0.030239)	0.839502 (0.372703)	0.752730 (0.294915)
6	0.380172 (0.259082)	-0.490878 (0.305990)	0.140110 (0.042981)	1.059191 (0.757432)	0.888866 (0.327070)
7	0.025551 (0.171274)	-0.135562 (0.219706)	0.092265 (0.027909)	1.023281 (0.437914)	0.631428 (0.280542)
8	0.196530 (0.184080)	-0.601285 (0.239300)	0.135118 (0.044037)	0.140686 (0.523497)	0.393298 (0.287358)
9	0.032224 (0.171682)	-0.012856 (0.231807)	0.084996 (0.023875)	0.813722 (0.384107)	0.795865 (0.232311)
10	0.184698 (0.178312)	-1.070692 (0.287361)	0.179045 (0.034929)	0.886349 (0.398575)	0.577111 (0.302879)

$$\boldsymbol{R}_x = \begin{bmatrix} 1 & 0.207371 & 0.088933 & 0.281805 & 0.239528 \\ 0.207371 & 1 & 0.436783 & 0.392573 & -0.113970 \\ 0.088933 & 0.436783 & 1 & 0.197859 & -0.042160 \\ 0.281805 & 0.392573 & 0.197859 & 1 & 0.036242 \\ 0.239528 & -0.113970 & -0.042160 & 0.036242 & 1 \end{bmatrix}$$

constructed a "working" or "imputed" correlation matrix, equal to:

$$\hat{\boldsymbol{R}}_x = \begin{bmatrix} 1 & 0.5 & 0 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 1 & 0.25 & 0 \\ 0.5 & 0.5 & 0.25 & 1 & 0 \\ 0.25 & 0 & 0 & 0 & 1 \end{bmatrix}$$

This matrix is very simplified, in the sense that large correlations were rounded to 0.25 or 0.5, whereas the smaller ones (which are also the statistically non-significant), were set to zero. In real life applications, such a matrix could have been observed, for instance, in one or more of the included studies, or alternatively, it could have been compiled by collecting pairwise correlations concerning the variables at hand from the literature. Of course, in many applications the obtained matrix would be closer to the actual one, but we deliberately used such a crude approximation in order to simulate a condition in which only vague prior knowledge is available (that is, that two variables are positively correlated or not).

The results of this sensitivity analysis are listed in **Table 2**. In the table, we also list the results of the regression on the pooled dataset. For reasons of completeness, we also present the results obtained by the so-called meta-analysis of Individual Patients Data (IPD), in which we perform a stratified (by "study") regression analysis with a linear mixed model with random coefficients for the independent variables [37].

Even though the interpretation of the results did not change in nearly all analyses, some useful conclusions can be drawn. First of all, four out of the five variables have large and significant effects on glucose (triceps skin fold thickness, insulin, BMI and age) whereas DBP show a negligible (non-significant) association. Most of the methods corroborate to thus, with the exception of the method of Riley which produces a marginally not-significant association for triceps skin fold thickness as well. As expected, the summary meta-analysis using the actual correlation matrix and the meta-analysis using IPD, yield similar even though not identical estimates. Concerning

Table 2. The estimates for the meta-analysis on $k = 10$ artificially generated "studies", obtained using the different methods. The regression coefficients for each subset (s) correspond to diastolic blood pressure (β_1), triceps skin fold thickness (β_2), 2-hour serum insulin (β_3), body mass index (β_4) and age (β_5). For the explanation of the methods, see the main text.

method	β_1 (*se*)	β_2 (*se*)	β_3 (*se*)	β_4 (*se*)	β_5 (*se*)
Ordinary linear regression on the pooled dataset ($n = 768$ subjects)	0.067157 (0.057144)	−0.319836 (0.077137)	0.102682 (0.009834)	0.771217 (0.144281)	0.664185 (0.090638)
Meta-analysis ($k = 10$ studies), using the actual variance-covariance matrix	0.077151 (0.066722)	−0.377597 (0.1288864)	0.108197 (0.012404)	0.818003 (0.156274)	0.657785 (0.118143)
Meta-analysis ($k = 10$ studies), using the actual variance estimates and assuming zero correlation	0.089251 (0.071367)	−0.374107 (0.131523)	0.110964 (0.012804)	0.819255 (0.156134)	0.657269 (0.114043)
Meta-analysis ($k = 10$ studies), using the actual variance estimates and the alternative model of Riley	0.024612 (0.055657)	−0.206025 (0.107372)	0.095521 (0.007975)	0.880839 (0.155768)	0.718012 (0.141181)
Meta-analysis ($k = 10$ studies), using the actual variance estimates and assuming a plausible correlation	0.071282 (0.065038)	−0.346298 (0.124092)	0.107940 (0.011789)	0.826325 (0.160682)	0.667961 (0.121064)
Meta-analysis ($k = 10$ studies), using a random coefficient model and Individual Patients Data (IPD)	0.078416 (0.064406)	−0.365146 (0.103916)	0.105337 (0.010688)	0.801249 (0.142996)	0.631803 (0.098856)

the other three approaches for summary data meta-analysis, the method that we proposed here, using the "working" correlation matrix, produces the results that resemble closely the ones obtained by using the actual correlation matrix of each "study". This is true for both the regression estimates and their standard errors. The naive method of assuming zero correlation and the method of Riley, produced slightly biased estimates and standard errors, which especially in the case of Riley's method yield a non-significant effect for one of the variables (triceps skin fold thickness). This can be explained, since the regression estimates for triceps skin fold thickness β_2 had the largest variability between studies and given that the method of Riley cannot differentiate the sources of variability, inflates this way the overall estimate of the variance of the particular coefficient. The dataset and the source code are given at http://www.compgen.org/tools/regression.

The source code that we provide, presents an easily applied and fast method for calculating the covariance matrix of the regression coefficients given the correlation matrix of the explanatory variables. We applied this method in two important problems, namely in the meta-analysis of regression coefficients and in synthesis analysis, with very encouraging results. Since the expression is mathematically equivalent to the already known expressions, when the correlations are the actual correlations of the sample the results are identical. However, even in the case where the actual correlations are not known from the sample, these can be imputed using data from the literature. In this case, as one would expect, the method is very robust to modest deviations from the actual values. Our results, build upon the earlier works of Riley and coworkers and Wu and Becker, and demonstrate the usefulness of the method. Thus, we knew that by ignoring the within studies correlation may result in biased estimates for the variance of the effect size, and that the alternative model may be useful in several circumstances. Now, we have an ever better approximation that can be used in order to obtain better results. The idea of calculating the correlation of estimates using the pairwise correlation of the variablesinvolved, has already being presented in a general meta-analysis setting [22] [23], and thus, we expect that this method can be useful both to meta-analysis of regression coefficients and to synthesis analysis.

When we reconstruct the correlation matrix using data from the literature, two things need to be addressed. First, we may encounter the problem of a non-positive definite covariance matrix [38]. The chance of this happening increases with the number of variables included and with increasing correlations among them. When two variables are highly correlated (correlation > 0.99), a simple solution would be to exclude one of them from the model. In all other cases, in order to overcome the problem, the most reasonable solution would be to transform the non-positive definite covariance matrix into positive definite. For this, we can use a simple heuristic consisting of adding the negative of the smallest eigenvalue (which will be negative) plus a small constant (10^{-7}) to the diagonal elements of the covariance matrix, or some other among the correction techniques proposed in the literature [38]-[40]. The second thing to remind, is that when we have multiple sources of evidence concerning a particular correlation, or for the whole correlation matrix, then, the obvious solution would be to pool them using appropriate meta-analysis methods. Methods for pooling correlation coefficients are known for years, but it will be advantageous, when possible, to pool the whole correlation matrix using a multivariate technique that properly takes their own covariances into account [41]-[44].

4. Conclusions

In this work, we derive an analytical expression for the covariance matrix of the regression coefficients in a multiple linear regression model. In contrast to the well-known expressions which make use of the cross-product matrix $\mathbf{X}^T\mathbf{X}$, we express the covariance matrix of the regression coefficients directly in terms of covariance matrix of the explanatory variables. This is very important since the covariance matrix of the explanatory variables can be easily obtained or imputed using data from the literature, without requiring access to individual data. In particular, we show that the covariance matrix of the regression coefficients can be calculated using the matrix of the partial correlation coefficients of the explanatory variables, which in turn can be calculated easily from the correlation matrix of the explanatory variables.

The estimate proposed in this work can be useful in several applications. As we already noted, meta-analysis of regression coefficients is increasingly being used in several applications both in the biological [6] [7] [45] as well as in the social sciences [8]-[11]. Thus, the estimate proposed here, coupled with the advances in multivariate meta-analysis software, can facilitate further the use of the method. Some other, more advanced techniques have also been proposed for synthesizing regression coefficients, especially when the studies are included in the meta-analysis evaluate different set of explanatory variables [46] [47]. However, these techniques require spe-

cialised software or user-written code, whereas the traditional approach mentioned here can be fitted using standard software for multivariate meta-analysis. Finally, the influence of the omitted variables (*i.e.* the variables that are not measured in some of the included studies), can be evaluated and adjusted for using multivariate meta-regression, simply by adding an indicator variable for each of the omitted covariates. We believe that such an approach will be efficient and easily used.

The method proposed here, can also greatly increase the usability of the standard synthesis analysis method. For instance, such methods can be used for constructing multivariate prognostic models using the univariate associations. Of particular importance is the ability to incorporate published univariable associations in diagnostic and prognostic models [14], or the ability to adjust the results of an individual data analysis, for another recently discovered factor, using estimates from the literature [14] [48].

Other potential applications can be found in the social sciences, where statistical methods for comparing regression coefficients between models [49] are needed, especially in the study of mediation models, such as in the case of psychology [50]. Moreover, as we showed in the manuscript, the method is already available for use also with the standardized regression coefficients (b). Even though the use of standardized regression coefficients in epidemiology has been the subject of debate [9] [51] [52], they are routinely used in the social sciences [53] and they become popular in genetics with genome-wide association studies [54]-[56]. Thus, we believe that the method can be useful also in this respect.

The assumptions, on which the method is based, need also to be discussed. For the derivation we assume that the dependent and the independent variables are jointly multivariate normally distributed. This is one of the two main approaches for formulating a regression problem (the other is the approach that assumes that the independent variables are fixed by design). Even though the two approaches are conceptually very different, it is well known that concerning the estimation of the regression parameters (the coefficients and their variance), they yield exactly the same results. The assumption of multivariate normality is more stringent, but it yields an optimal predictor among all choices, rather than merely among linear predictors. Practically, since the estimators are identical, this means that we can use the expressions derived here, even in the case of binary independent variables and in any case the results are identical with the ones produced by any standard linear regression software. We need to mention at this point that the method is developed in [13], which as the authors claimed does not make the assumption of normality, yields estimates for the regression coefficients that differ from the ones produced by standard regression packages.

When it comes to binary dependent variables however, the situation is more complicated. The method can also be used, after appropriate transformations, for estimating the parameters of such models (*i.e.* logistic regression). Several similar methods have been proposed in the literature [57] [58], but they are all based on the method of Cornfield [59], which is approximate and produces biased estimates [60]-[62]. This fact, along with some other fundamental differences between the linear model and the logistic regression model [63] [64], rings the bell for the use of such methods, and makes imperative the need for new accurate methods for binary data.

Acknowledgements

This work is part of the project “IntDaMuS: Integration of Data from Multiple Sources” which is implemented under the “ARISTEIA II”. Action of the “OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING” and is co-funded by the European Social Fund (ESF) and National Resources.

References

[1] Platt, R.W. (1998) ANOVA, *t* Tests, and Linear Regression. *Injury Prevention*, **4**, 52-53. http://dx.doi.org/10.1136/ip.4.1.52

[2] Vickers, A.J. (2005) Analysis of Variance Is Easily Misapplied in the Analysis of Randomized Trials: A Critique and Discussion of Alternative Statistical Approaches. *Psychosomatic Medicine*, **67**, 652-655. http://dx.doi.org/10.1097/01.psy.0000172624.52957.a8

[3] Becker, B.J. and Wu, M.J. (2007) The Synthesis of Regression Slopes in Meta-Analysis. *Statistical Science*, **22**, 414-429. http://dx.doi.org/10.1214/07-STS243

[4] Mavridis, D. and Salanti, G. (2013) A Practical Introduction to Multivariate Meta-Analysis. *Statistical Methods in Medical Research*, **22**, 133-158. http://dx.doi.org/10.1177/0962280211432219

[5] van Houwelingen, H.C., Arends, L.R. and Stijnen, T. (2002) Advanced Methods in Meta-Analysis: Multivariate Ap-

proach and Meta-Regression. *Statistics in Medicine*, **21**, 589-624. http://dx.doi.org/10.1002/sim.1040

[6] Manning, A.K., LaValley, M., Liu, C.T., Rice, K., An, P., Liu, Y., Miljkovic, I., Rasmussen-Torvik, L., Harris, T.B., Province, M.A., Borecki, I.B., Florez, J.C., Meigs, J.B., Cupples, L.A. and Dupuis, J. (2011) Meta-Analysis of Gene-Environment Interaction: Joint Estimation of SNP and SNP x Environment Regression Coefficients. *Genetic Epidemiology*, **35**, 11-18. http://dx.doi.org/10.1002/gepi.20546

[7] Paul, P.A., Lipps, P.E. and Madden, L.V. (2006) Meta-Analysis of Regression Coefficients for the Relationship between Fusarium Head Blight and Deoxynivalenol Content of Wheat. *Phytopathology*, **96**, 951-961. http://dx.doi.org/10.1094/PHYTO-96-0951

[8] Rose, A.K. and Stanley, T.D. (2005) A Meta-Analysis of the Effect of Common Currencies on International Trade. *Journal of Economic Surveys*, **19**, 347-365. http://dx.doi.org/10.1111/j.0950-0804.2005.00251.x

[9] Peterson, R.A. and Brown, S.P. (2005) On the Use of Beta Coefficients in Meta-Analysis. *Journal of Applied Psychology*, **90**, 175-181. http://dx.doi.org/10.1037/0021-9010.90.1.175

[10] Crouch, G.I. (1995) A Meta-Analysis of Tourism Demand. *Annals of Tourism Research*, **22**, 103-118. http://dx.doi.org/10.1016/0160-7383(94)00054-V

[11] Aloe, A.M. and Becker, B.J. (2011) Advances in Combining Regression Results in Meta-Analysis. In: Williams, M. and Vogt, W.P., Eds., *The SAGE Handbook of Innovation in Social Research Methods*, SAGE, London, 331-352. http://dx.doi.org/10.4135/9781446268261.n20

[12] Samsa, G., Hu, G. and Root, M. (2005) Combining Information from Multiple Data Sources to Create Multivariable Risk Models: Illustration and Preliminary Assessment of a New Method. *Journal of Biomedicine and Biotechnology*, **2005**, 113-123. http://dx.doi.org/10.1155/JBB.2005.113

[13] Zhou, X.H., Hu, N., Hu, G. and Root, M. (2009) Synthesis Analysis of Regression Models with a Continuous Outcome. *Statistics in Medicine*, **28**, 1620-1635. http://dx.doi.org/10.1002/sim.3563

[14] Debray, T.P., Koffijberg, H., Lu, D., Vergouwe, Y., Steyerberg, E.W. and Moons, K.G. (2012) Incorporating Published Univariable Associations in Diagnostic and Prognostic Modeling. *BMC Medical Research Methodology*, **12**, 121. http://dx.doi.org/10.1186/1471-2288-12-121

[15] Noble, D., Mathur, R., Dent, T., Meads, C. and Greenhalgh, T. (2011) Risk Models and Scores for Type 2 Diabetes: Systematic Review. *BMJ*, **343**, d7163. http://dx.doi.org/10.1136/bmj.d7163

[16] Moons, K.G., Kengne, A.P., Grobbee, D.E., Royston, P., Vergouwe, Y., Altman, D.G. and Woodward, M. (2012) Risk Prediction Models: II. External Validation, Model Updating, and Impact Assessment. *Heart*, **98**, 691-698. http://dx.doi.org/10.1136/heartjnl-2011-301247

[17] van Dieren, S., Beulens, J.W., Kengne, A.P., Peelen, L.M., Rutten, G.E., Woodward, M., van der Schouw, Y.T. and Moons, K.G. (2012) Prediction Models for the Risk of Cardiovascular Disease in Patients with Type 2 Diabetes: A Systematic Review. *Heart*, **98**, 360-369. http://dx.doi.org/10.1136/heartjnl-2011-300734

[18] Jackson, D., Riley, R. and White, I.R. (2011) Multivariate Meta-Analysis: Potential and Promise. *Statistics in Medicine*, **30**, 2481-2498. http://dx.doi.org/10.1002/sim.4172

[19] Riley, R.D., Abrams, K.R., Lambert, P.C., Sutton, A.J. and Thompson, J.R. (2007) An Evaluation of Bivariate Random-Effects Meta-Analysis for the Joint Synthesis of Two Correlated Outcomes. *Statistics in Medicine*, **26**, 78-97. http://dx.doi.org/10.1002/sim.2524

[20] Riley, R.D., Abrams, K.R., Sutton, A.J., Lambert, P.C. and Thompson, J.R. (2007) Bivariate Random-Effects Meta-Analysis and the Estimation of Between-Study Correlation. *BMC Medical Research Methodology*, **7**, 3. http://dx.doi.org/10.1186/1471-2288-7-3

[21] Riley, R.D., Thompson, J.R. and Abrams, K.R. (2008) An Alternative Model for Bivariate Random-Effects Meta-Analysis When the Within-Study Correlations Are Unknown. *Biostatistics*, **9**, 172-186. http://dx.doi.org/10.1093/biostatistics/kxm023

[22] Bagos, P.G. (2012) On the Covariance of Two Correlated Log-Odds Ratios. *Statistics in Medicine*, **31**, 1418-1431. http://dx.doi.org/10.1002/sim.4474

[23] Wei, Y. and Higgins, J.P. (2013) Estimating Within-Study Covariances in Multivariate Meta-Analysis with Multiple Outcomes. *Statistics in Medicine*, **32**, 1191-1205.

[24] Green, W. (2008) Econometric Analysis. 6th Edition, Prentice Hall, Englewood Cliffs.

[25] Hsieh, F.Y., Bloch, D.A. and Larsen, M.D. (1998) A Simple Method of Sample Size Calculation for Linear and Logistic Regression. *Statistics in Medicine*, **17**, 1623-1634. http://dx.doi.org/10.1002/(SICI)1097-0258(19980730)17:14<1623::AID-SIM871>3.0.CO;2-S

[26] O'Brien, R. (2007) A Caution regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, **41**, 673-690. http://dx.doi.org/10.1007/s11135-006-9018-6

[27] Johnson, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis. 6th Edition, Pearson Prentice Hall, Upper Saddle River.

[28] Dwyer, P.S. (1940) The Evaluation of Multiple and Partial Correlation Coefficients from the Factorial Matrix. *Psychometrika*, **5**, 211-232. http://dx.doi.org/10.1007/BF02288567

[29] Stapleton, J.H. (1995) Linear Statistical Models. John Wiley & Sons, Inc., Hoboken. http://dx.doi.org/10.1002/9780470316924

[30] Rencher, A.C. (1995) Methods of Multivariate Analysis. John Wiley & Sons, Inc., New York.

[31] Weisberg, S. (2005) Applied Linear Regression. 3rd Edition, Wiley/Interscience, Hoboken. http://dx.doi.org/10.1002/0471704091

[32] Timm, N.H. (2002) Applied Multivariate Analysis. Springer-Verlag Inc., New York.

[33] Seber, G.A.F. and Lee, A.J. (2003) Linear Regression Analysis. John Wiley & Sons, Inc., Hoboken. http://dx.doi.org/10.1002/9780471722199

[34] Bache, K. and Lichman, M. (2015) UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[35] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C. and Johannes, R.S. (1988) Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, Orlando, 7-11 November, 261-265.

[36] White, I.R. (2009) Multivariate Random-Effects Meta-Analysis. *Stata Journal*, **9**, 40-56.

[37] Higgins, J.P., Whitehead, A., Turner, R.M., Omar, R.Z. and Thompson, S.G. (2001) Meta-Analysis of Continuous Outcome Data from Individual Patients. *Statistics in Medicine*, **20**, 2219-2241. http://dx.doi.org/10.1002/sim.918

[38] Schwertman, N.C. and Allen, D.M. (1979) Smoothing an Indefinite Variance-Covariance Matrix. *Journal of Statistical Computation and Simulation*, **9**, 183-194. http://dx.doi.org/10.1080/00949657908810316

[39] Rebonato, R. and Jäckel, P. (1999) The Most General Methodology to Create a Valid Correlation Matrix for Risk Management and Option Pricing Purposes. *Journal of Risk*, **2**, 17-28.

[40] Higham, N.J. (2002) Computing the Nearest Correlation Matrix—A Problem from Finance. *IMA Journal of Numerical Analysis*, **22**, 329-343. http://dx.doi.org/10.1093/imanum/22.3.329

[41] Field, A.P. (2001) Meta-Analysis of Correlation Coefficients: A Monte Carlo Comparison of Fixed- and Random-Effects Methods. *Psychological Methods*, **6**, 161-180. http://dx.doi.org/10.1037/1082-989X.6.2.161

[42] Hafdahl, A.R. (2007) Combining Correlation Matrices: Simulation Analysis of Improved Fixed-Effects Methods. *Journal of Educational and Behavioral Statistics*, **32**, 180-205. http://dx.doi.org/10.3102/1076998606298041

[43] Hafdahl, A.R. and Williams, M.A. (2009) Meta-Analysis of Correlations Revisited: Attempted Replication and Extension of Field's (2001) Simulation Studies. *Psychological Methods*, **14**, 24-42. http://dx.doi.org/10.1037/a0014697

[44] Prevost, A.T., Mason, D., Griffin, S., Kinmonth, A.L., Sutton, S. and Spiegelhalter, D. (2007) Allowing for Correlations between Correlations in Random-Effects Meta-Analysis of Correlation Matrices. *Psychological Methods*, **12**, 434-450. http://dx.doi.org/10.1037/1082-989X.12.4.434

[45] Debray, T.P., Koffijberg, H., Nieboer, D., Vergouwe, Y., Steyerberg, E.W. and Moons, K.G. (2014) Meta-Analysis and Aggregation of Multiple Published Prediction Models. *Statistics in Medicine*, **33**, 2341-2362. http://dx.doi.org/10.1002/sim.6080

[46] Wu, M.J. and Becker, B.J. (2013) Synthesizing Regression Results: A Factored Likelihood Method. *Research Synthesis Methods*, **4**, 127-143. http://dx.doi.org/10.1002/jrsm.1063

[47] Dominici, F., Parmigiani, G., Reckhow, K.H. and Wolper, R.L. (1997) Combining Information from Related Regressions. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 313-332. http://dx.doi.org/10.2307/1400448

[48] Steyerberg, E.W., Eijkemans, M.J., Van Houwelingen, J.C., Lee, K.L. and Habbema, J.D. (2000) Prognostic Models Based on Literature and Individual Patient Data in Logistic Regression Analysis. *Statistics in Medicine*, **19**, 141-160. http://dx.doi.org/10.1002/(SICI)1097-0258(20000130)19:2<141::AID-SIM334>3.0.CO;2-O

[49] Clogg, C.C., Petkova, E. and Haritou, A. (1995) Statistical Methods for Comparing Regression Coefficients between Models. *American Journal of Sociology*, **10**, 1261-1293. http://dx.doi.org/10.1086/230638

[50] Tofighi, D., Mackinnon, D.P. and Yoon, M. (2009) Covariances between Regression Coefficient Estimates in a Single Mediator Model. *British Journal of Mathematical and Statistical Psychology*, **62**, 457-484.

[51] Greenland, S., Schlesselman, J.J. and Criqui, M.H. (1986) The Fallacy of Employing Standardized Regression Coefficients and Correlations as Measures of Effect. *American Journal of Epidemiology*, **123**, 203-208.

[52] Greenland, S., Maclure, M., Schlesselman, J.J., Poole, C. and Morgenstern, H. (1991) Standardized Regression Coeffi-

cients: A Further Critique and Review of Some Alternatives. *Epidemiology*, **2**, 387-392. http://dx.doi.org/10.1097/00001648-199109000-00015

[53] Cheung, M.W. (2009) Comparison of Methods for Constructing Confidence Intervals of Standardized Indirect Effects. *Behavior Research Methods*, **41**, 425-438. http://dx.doi.org/10.3758/BRM.41.2.425

[54] Begum, F., Ghosh, D., Tseng, G.C. and Feingold, E. (2012) Comprehensive Literature Review and Statistical Considerations for GWAS Meta-Analysis. *Nucleic Acids Research*, **40**, 3777-3784. http://dx.doi.org/10.1093/nar/gkr1255

[55] Evangelou, E. and Ioannidis, J.P. (2013) Meta-Analysis Methods for Genome-Wide Association Studies and Beyond. *Nature Reviews Genetics*, **14**, 379-389. http://dx.doi.org/10.1038/nrg3472

[56] Cantor, R.M., Lange, K. and Sinsheimer, J.S. (2010) Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *American Journal of Human Genetics*, **86**, 6-22. http://dx.doi.org/10.1016/j.ajhg.2009.11.017

[57] Sheng, E., Zhou, X.H., Chen, H., Hu, G. and Duncan, A. (2014) A New Synthesis Analysis Method for Building Logistic Regression Prediction Models. *Statistics in Medicine*, **33**, 2567-2576. http://dx.doi.org/10.1002/sim.6125

[58] Chang, B.-H., Liopsitz, S. and Waternaux, C. (2000) Logistic Regression in Meta-Analysis Using Aggregate Data. *Journal of Applied Statistics*, **27**, 411-424. http://dx.doi.org/10.1080/02664760050003605

[59] Cornfield, J. (1962) Joint Dependence of Risk of Coronary Heart Disease on Serum Cholesterol and Systolic Blood Pressure: A Discriminant Function Analysis. *Federation Proceedings*, **21**, 58-61.

[60] Halperin, M., Blackwelder, W.C. and Verter, J.I. (1971) Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches. *Journal of Chronic Diseases*, **24**, 125-158. http://dx.doi.org/10.1016/0021-9681(71)90106-8

[61] Hosmer, T., Hosmer, D. and Fisher, L. (1983) A Comparison of the Maximum Likelihood and Discriminant Function Estimators of the Coefficients of the Logistic Regression Model for Mixed Continuous and Discrete Variables. *Communications in Statistics—Simulation and Computation*, **12**, 23-43. http://dx.doi.org/10.1080/03610918308812298

[62] Press, S.J. and Wilson, S. (1978) Choosing between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, **73**, 699-705. http://dx.doi.org/10.1080/01621459.1978.10480080

[63] Xing, G. and Xing, C. (2010) Adjusting for Covariates in Logistic Regression Models. *Genetic Epidemiology*, **34**, 769-771; Author Reply 772. http://dx.doi.org/10.1002/gepi.20526

[64] Robinson, L.D. and Jewell, N.P. (1991) Some Surprising Results about Covariate Adjustment in Logistic Regression Models. *International Statistical Review*, **59**, 227-240. http://dx.doi.org/10.2307/1403444

Appendix A

Consider the $p \times p$ diagonal matrix $\boldsymbol{V}_x$ such that $\boldsymbol{V}_x = \text{diag}\left(s_{x_1}, s_{x_2}, \cdots, s_{x_p}\right)$. From Equation (4) and Equation (16), it is obvious that

$$\boldsymbol{\Sigma}_{22} = \boldsymbol{V}_x \boldsymbol{R}_x \boldsymbol{V}_x, \tag{A.1}$$

which implies:

$$\boldsymbol{R}_x^{-1} = \boldsymbol{V}_x \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{V}_x \tag{A.2}$$

Reminding that for each $i = 1, 2, \cdots, p$ holds

$$a_{1i} = \frac{s_{x_i y}}{s_{x_i}^2} \tag{A.3}$$

Using Equations (5), (A.3), (A.2), (16) and the Hadamard product $\odot$, we can write:

$$\begin{aligned} \boldsymbol{R}_x^{-1}\left(\boldsymbol{A} \odot \boldsymbol{S}\right) &= \boldsymbol{R}_x^{-1}\left(\begin{bmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1p} \end{bmatrix} \odot \begin{bmatrix} s_{x_1} \\ s_{x_2} \\ \vdots \\ s_{x_p} \end{bmatrix}\right) = \boldsymbol{R}_p^{-1}\left(\begin{bmatrix} s_{x_1 y} / s_{x_1}^2 \\ s_{x_2 y} / s_{x_2}^2 \\ \vdots \\ s_{x_p y} / s_{x_p}^2 \end{bmatrix} \odot \begin{bmatrix} s_{x_1} \\ s_{x_2} \\ \vdots \\ s_{x_p} \end{bmatrix}\right) = \boldsymbol{V}_x \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{V}_x \begin{bmatrix} s_{yx_1} / s_{x_1} \\ s_{yx_2} / s_{x_2} \\ \vdots \\ s_{yx_p} / s_{x_p} \end{bmatrix} = \boldsymbol{V}_x \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{V}_x \begin{bmatrix} s_{x_1 y} / s_{x_1} \\ s_{x_2 y} / s_{x_2} \\ \vdots \\ s_{x_p y} / s_{x_p} \end{bmatrix} \\ &= \boldsymbol{V}_x \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{V}_x \boldsymbol{V}_x^{-1} \begin{bmatrix} s_{x_1 y} \\ s_{x_2 y} \\ \vdots \\ s_{x_p y} \end{bmatrix} = \boldsymbol{V}_x \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \end{aligned} \tag{A.4}$$

Denoting

$$\boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix} \tag{A.5}$$

Equation (A.4) yields:

$$\boldsymbol{R}_x^{-1}\left(\boldsymbol{A} \odot \boldsymbol{S}\right) = \boldsymbol{V}_x \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} = \boldsymbol{V}_x \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix} = \begin{bmatrix} s_{x_1} u_1 \\ s_{x_2} u_2 \\ \vdots \\ s_{x_p} u_p \end{bmatrix} \tag{A.6}$$

Finally, denoting

$$\frac{1}{\boldsymbol{S}} = \begin{bmatrix} 1/s_{x_1} \\ 1/s_{x_2} \\ \vdots \\ 1/s_{x_p} \end{bmatrix}$$

and combining Equations (A.6) and (A.5) we derive:

$$\frac{\boldsymbol{R}_x^{-1}\left(\boldsymbol{A} \odot \boldsymbol{S}\right)}{\boldsymbol{S}} = \left(\boldsymbol{R}_x^{-1}\left(\boldsymbol{A} \odot \boldsymbol{S}\right)\right) \odot \frac{1}{\boldsymbol{S}} = \begin{bmatrix} s_{x_1} u_1 \\ s_{x_2} u_2 \\ \vdots \\ s_{x_p} u_p \end{bmatrix} \odot \begin{bmatrix} 1/s_{x_1} \\ 1/s_{x_2} \\ \vdots \\ 1/s_{x_p} \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix} = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$$

Appendix B

Consider the $p\times 1$ matrix of the standardized regression coefficients $\boldsymbol{b}$, the well known correlation matrices $\boldsymbol{R}_x$ and $\boldsymbol{R}_{yx}$

$$\boldsymbol{b}=\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix},\quad \boldsymbol{R}_x=\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix},\quad \boldsymbol{R}_{yx}=\begin{bmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{py} \end{bmatrix}$$

and, the $p\times p$ diagonal matrix $\boldsymbol{V}_x$ such that. As in Appendix A, Equation (A.1) yields:

$$\boldsymbol{R}_x^{-1}=\boldsymbol{V}_x\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{V}_x \tag{B.1}$$

Using Equation (B.1), Equation (16) and Equation (17), we derive:

$$\begin{aligned}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{yx}&=\boldsymbol{V}_x\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{V}_x\boldsymbol{R}_{yx}=\boldsymbol{V}_x\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{V}_x\begin{bmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{py} \end{bmatrix}=\boldsymbol{V}_x\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{V}_x\begin{bmatrix} s_{x_1y}/\left(s_{x_1}s_y\right) \\ s_{x_2y}/\left(s_{x_2}s_y\right) \\ \vdots \\ s_{x_py}/\left(s_{x_p}s_y\right) \end{bmatrix}=\boldsymbol{V}_x\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{V}_x\boldsymbol{V}_x^{-1}\begin{bmatrix} s_{x_1y}/s_y \\ s_{x_2y}/s_y \\ \vdots \\ s_{x_py}/s_y \end{bmatrix} \\ &=\frac{1}{s_y}\boldsymbol{V}_x\boldsymbol{\Sigma}_{22}^{-1}\begin{bmatrix} s_{x_1y} \\ s_{x_2y} \\ \vdots \\ s_{x_py} \end{bmatrix}=\frac{1}{s_y}\boldsymbol{V}_x\boldsymbol{\Sigma}_{22}^{-1}\begin{bmatrix} s_{yx_1} \\ s_{yx_2} \\ \vdots \\ s_{yx_p} \end{bmatrix}=\frac{1}{s_y}\boldsymbol{V}_x\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}=\frac{1}{s_y}\boldsymbol{V}_x\boldsymbol{\beta}\end{aligned} \tag{B.2}$$

Equation (B.2) follows:

$$\boldsymbol{\beta}=s_y\boldsymbol{V}_x^{-1}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{yx}$$

Appendix C

Let $\boldsymbol{X}_c$ be an $n\times p$ matrix, $\mathbf{1}_n$ be the $n\times 1$ matrix of 1 s, with

$$\boldsymbol{X}_c=\begin{bmatrix} x_1-\overline{x}_1\mathbf{1}_n & x_2-\overline{x}_2\mathbf{1}_n & \cdots & x_p-\overline{x}_p\mathbf{1}_n \end{bmatrix}=\begin{bmatrix} \tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_p \end{bmatrix}$$

It is well known that

$$\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c=\boldsymbol{V}_c\boldsymbol{R}_x\boldsymbol{V}_c, \tag{C.1}$$

where the $p\times p$ diagonal matrix $\boldsymbol{V}_c$ is define

$\boldsymbol{V}_c=\operatorname{diag}\left(\sqrt{\sum\tilde{x}_1^2},\sqrt{\sum\tilde{x}_2^2},\cdots,\sqrt{\sum\tilde{x}_p^2}\right)$ and $\boldsymbol{R}_x$ denotes the correlation matrix.

It is obvious that

$$\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c=\begin{bmatrix} \sum\tilde{x}_1^2 & \sum\tilde{x}_1\tilde{x}_2 & \cdots & \sum\tilde{x}_1\tilde{x}_p \\ \sum\tilde{x}_1\tilde{x}_2 & \sum\tilde{x}_2^2 & \cdots & \sum\tilde{x}_2\tilde{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \sum\tilde{x}_1\tilde{x}_p & \sum\tilde{x}_2\tilde{x}_p & \cdots & \sum\tilde{x}_p^2 \end{bmatrix}$$

and using the definition of the Pearson's correlation coefficient

$$r_{ij}=\frac{\sum\tilde{x}_i\tilde{x}_j}{\sqrt{\sum\tilde{x}_i^2}\sqrt{\sum\tilde{x}_j^2}} \tag{C.2}$$

for $i, j = 1, 2, \cdots, p$, the matrix $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$ is written as:

$$\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c = \begin{bmatrix} \sum \tilde{x}_1^2 & \sqrt{\sum \tilde{x}_1^2}\sqrt{\sum \tilde{x}_2^2}\, r_{12} & \cdots & \sqrt{\sum \tilde{x}_1^2}\sqrt{\sum \tilde{x}_p^2}\, r_{1p} \\ \sqrt{\sum \tilde{x}_1^2}\sqrt{\sum \tilde{x}_2^2}\, r_{12} & \sum \tilde{x}_2^2 & \cdots & \sqrt{\sum \tilde{x}_2^2}\sqrt{\sum \tilde{x}_p^2}\, r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\sum \tilde{x}_1^2}\sqrt{\sum \tilde{x}_p^2}\, r_{1p} & \sqrt{\sum \tilde{x}_2^2}\sqrt{\sum \tilde{x}_p^2}\, r_{2p} & \cdots & \sum \tilde{x}_p^2 \end{bmatrix} \tag{C.3}$$

Notice that from (C.1) arises

$$\det\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right) = \left(\det \boldsymbol{V}_c\right)^2 \det \boldsymbol{R}_x = \prod_{m=1}^{p}\left(\sum \tilde{x}_m^2\right)\det \boldsymbol{R}_x \tag{C.4}$$

Furthermore, for the matrix $\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \beta_2 & \cdots & \beta_p\end{bmatrix}^{\mathrm{T}}$

$$\begin{aligned} \operatorname{cov}(\boldsymbol{\beta}) &= \sigma^2\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)^{-1} \\ &= \frac{\sigma^2}{\det\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)}\operatorname{adj}\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right), \end{aligned} \tag{C.5}$$

where $\operatorname{adj}\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)$ denotes the $p \times p$ adjoint matrix of $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$, (see Applied Linear regression, Sanford Weisberg (2007), p. 57). Adjoint is defined the matrix, whose (j,i) th element is formulated as:

$$(-1)^{i+j}\det\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)_{ij} \tag{C.6}$$

where $\det\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)_{ij}$ denotes the determinant of the $(p-1)\times(p-1)$ submatrix of $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$ obtained by deleting the i-th row and j-th column of $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$. Remind that $\operatorname{adj}\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)$ is a symmetric matrix, since $\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)^{\mathrm{T}} = \boldsymbol{X}_c^{\mathrm{T}}\left(\boldsymbol{X}_c^{\mathrm{T}}\right)^{\mathrm{T}} = \boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$. Combining the above remark, Equation (C.6), the properties of determinant and $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$ by Equation (C.3), the (i,j) th element of $\operatorname{adj}\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)$ can be written as:

$$\begin{aligned} &(-1)^{i+j}\det\left(\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c\right)_{ij} \\ &= (-1)^{i+j}\det\begin{bmatrix} \sum \tilde{x}_1^2 & \cdots & \sqrt{\sum \tilde{x}_1^2 \sum \tilde{x}_{j-1}^2}\, r_{1(j-1)} & \sqrt{\sum \tilde{x}_1^2 \sum \tilde{x}_{j+1}^2}\, r_{1(j+1)} & \cdots & \sqrt{\sum \tilde{x}_1^2 \sum \tilde{x}_p^2}\, r_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ \sqrt{\sum \tilde{x}_1^2 \sum \tilde{x}_{(i-1)}^2}\, r_{1(i-1)} & \cdots & \sqrt{\sum \tilde{x}_{(i-1)}^2 \sum \tilde{x}_{(j-1)}^2}\, r_{(i-1)(j-1)} & \sqrt{\sum \tilde{x}_{(i-1)}^2 \sum \tilde{x}_{(j+1)}^2}\, r_{(i-1)(j+1)} & \cdots & \sqrt{\sum \tilde{x}_{i-1}^2 \sum \tilde{x}_p^2}\, r_{(i-1)p} \\ \sqrt{\sum \tilde{x}_1^2 \sum \tilde{x}_{(i+1)}^2}\, r_{1(i+1)} & \cdots & \sqrt{\sum \tilde{x}_{(i+1)}^2 \sum \tilde{x}_{(j-1)}^2}\, r_{(i+1)(j-1)} & \sqrt{\sum x_{(i+1)}^2 \sum \tilde{x}_{(j+1)}^2}\, r_{(i+1)(j+1)} & \cdots & \sqrt{\sum \tilde{x}_{i+1}^2 \sum \tilde{x}_p^2}\, r_{(i+1)p} \\ \vdots & & \vdots & \vdots & & \vdots \\ \sqrt{\sum \tilde{x}_1^2 \sum \tilde{x}_p^2}\, r_{1p} & \cdots & \sqrt{\sum \tilde{x}_p^2 \sum \tilde{x}_{(j-1)}^2}\, r_{p(j-1)} & \sqrt{\sum \tilde{x}_p^2 \sum \tilde{x}_{(j+1)}^2}\, r_{p(j+1)} & \cdots & \sum \tilde{x}_p^2 \end{bmatrix} \\ &= (-1)^{i+j}\left(\prod_{\substack{m=1 \\ m\neq j}}^{p}\sqrt{\sum \tilde{x}_m^2}\right)\det\begin{bmatrix} \sqrt{\sum \tilde{x}_1^2} & \cdots & \sqrt{\sum \tilde{x}_1^2}\, r_{1(j-1)} & \sqrt{\sum \tilde{x}_1^2}\, r_{1(j+1)} & \cdots & \sqrt{\sum \tilde{x}_1^2}\, r_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ \sqrt{\sum \tilde{x}_{(i-1)}^2}\, r_{1(i-1)} & \cdots & \sqrt{\sum \tilde{x}_{(i-1)}^2}\, r_{(i-1)(j-1)} & \sqrt{\sum \tilde{x}_{(i-1)}^2}\, r_{(i-1)(j+1)} & \cdots & \sqrt{\sum \tilde{x}_{i-1}^2}\, r_{(i-1)p} \\ \sqrt{\sum \tilde{x}_{(i+1)}^2}\, r_{1(i+1)} & \cdots & \sqrt{\sum \tilde{x}_{(i+1)}^2}\, r_{(i+1)(j-1)} & \sqrt{\sum \tilde{x}_{(i+1)}^2}\, r_{(i+1)(j+1)} & \cdots & \sqrt{\sum \tilde{x}_{i+1}^2}\, r_{(i+1)p} \\ \vdots & & \vdots & \vdots & & \vdots \\ \sqrt{\sum \tilde{x}_p^2}\, r_{1p} & \cdots & \sqrt{\sum \tilde{x}_p^2}\, r_{p(j-1)} & \sqrt{\sum \tilde{x}_p^2}\, r_{p(j+1)} & \cdots & \sqrt{\sum \tilde{x}_p^2} \end{bmatrix} \end{aligned}$$

$$= (-1)^{i+j} \left(\prod_{\substack{m=1 \\ m \neq j}}^{p} \sqrt{\sum x_m^2} \right) \left(\prod_{\substack{m=1 \\ m \neq i}}^{p} \sqrt{\sum x_m^2} \right) \det \begin{bmatrix} 1 & \cdots & r_{1(j-1)} & r_{1(j+1)} & \cdots & r_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ r_{1(i-1)} & \cdots & r_{(i-1)(j-1)} & r_{(i-1)(j+1)} & \cdots & r_{(i-1)p} \\ r_{1(i+1)} & \cdots & r_{(i+1)(j-1)} & r_{(i+1)(j+1)} & \cdots & r_{(i+1)p} \\ \vdots & & \vdots & \vdots & & \vdots \\ r_{1p} & \cdots & r_{p(j-1)} & r_{p(j+1)} & \cdots & 1 \end{bmatrix}$$

$$= (-1)^{i+j} \Sigma_c \det \left(\boldsymbol{R}_x \right)_{ij}$$

Thus, we conclude

$$(-1)^{i+j} \det \left(\boldsymbol{X}_c^{\mathrm{T}} \boldsymbol{X}_c \right)_{ij} = (-1)^{i+j} \Sigma_c \det \left(\boldsymbol{R}_x \right)_{ij} \tag{C.7}$$

where

$$\Sigma_c = \sum \tilde{x}_1^2 \cdots \sum \tilde{x}_{i-1}^2 \sqrt{\sum \tilde{x}_i^2} \sum \tilde{x}_{i+1}^2 \cdots \sum \tilde{x}_{j-1}^2 \sqrt{\sum \tilde{x}_j^2} \sum \tilde{x}_{j+1}^2 \cdots \sum \tilde{x}_p^2 \tag{C.8}$$

In Equation (C.5) the (i, j)th element of $\operatorname{cov}(\boldsymbol{\beta})$ is denoted $\operatorname{cov}(\beta_i, \beta_j)$ and obtained by Equation (C.4), Equation (C.7) and Equation (C.8). In particular,

$$\begin{aligned} \operatorname{cov}(\beta_i, \beta_j) &= \frac{(-1)^{i+j} \sigma^2 \Sigma_c \det \left(\boldsymbol{R}_x \right)_{ij}}{\prod_{m=1}^{p} \left(\sum \tilde{x}_m^2 \right) \det \boldsymbol{R}_x} \\ &= \frac{(-1)^{i+j} \sigma^2 \sum \tilde{x}_1^2 \cdots \sum \tilde{x}_{i-1}^2 \sqrt{\sum \tilde{x}_i^2} \sum \tilde{x}_{i+1}^2 \cdots \sum \tilde{x}_{j-1}^2 \sqrt{\sum \tilde{x}_j^2} \sum \tilde{x}_{j+1}^2 \cdots \sum \tilde{x}_p^2 \det \left(\boldsymbol{R}_x \right)_{ij}}{\sum \tilde{x}_1^2 \cdots \sum \tilde{x}_{i-1}^2 \sum \tilde{x}_i^2 \sum \tilde{x}_{i+1}^2 \cdots \sum \tilde{x}_{j-1}^2 \sum \tilde{x}_j^2 \sum \tilde{x}_{j+1}^2 \cdots \sum \tilde{x}_p^2 \det \boldsymbol{R}_x} \\ &= \frac{(-1)^{i+j} \sigma^2 \det \left(\boldsymbol{R}_x \right)_{ij}}{\sqrt{\sum \tilde{x}_i^2} \sqrt{\sum \tilde{x}_j^2} \det \boldsymbol{R}_x} \end{aligned}$$

and using the definition of correlation r_{ij} by Equation (C.2) the above equation is written

$$\begin{aligned} \operatorname{cov}(\beta_i, \beta_j) &= \frac{(-1)^{i+j} \sigma^2 \det \left(\boldsymbol{R}_x \right)_{ij}}{\sqrt{\sum \tilde{x}_i^2} \sqrt{\sum \tilde{x}_j^2} \det \boldsymbol{R}_x} \\ &= \frac{(-1)^{i+j} \sigma^2 r_{ij} \det \left(\boldsymbol{R}_x \right)_{ij}}{\sum \tilde{x}_i \tilde{x}_j \det \boldsymbol{R}_x} \end{aligned} \tag{C.9}$$

$$= \frac{(-1)^{i+j} \sigma^2 r_{ij} \det \left(\boldsymbol{R}_x \right)_{ij}}{(n-1) s_{x_i x_j} \det \boldsymbol{R}_x} \tag{C.10}$$

The *i*-multiple correlation coefficient is denoted by R_i or $R_{i|1\cdots(i-1)((i+1)\cdots p}$ and given [28]

$$1 - R_i^2 = \frac{\det \boldsymbol{R}_x}{\det \left(\boldsymbol{R}_x \right)_{ii}}$$

whereby arises

$$\det \boldsymbol{R}_x = \left(1 - R_i^2 \right) \det \left(\boldsymbol{R}_x \right)_{ii} \tag{C.11}$$

Remind that in Equation (C.11) the correlation matrix $\boldsymbol{R}_x$ is a symmetric positive definite matrix, hence $\det \boldsymbol{R}_x > 0$ and $\det \left(\boldsymbol{R}_x \right)_{ii} > 0$, for every $i = 1, 2, \cdots, p$, as a main submatrix of $\boldsymbol{R}_x$. Since Equation (C.11) yields

$$\sqrt{\det \boldsymbol{R}_x} = \sqrt{\left(1-R_i^2\right)}\sqrt{\det\left(\boldsymbol{R}_x\right)_{ii}}$$

by the above equality for $i \neq j$ we can write:

$$\det \boldsymbol{R}_x = \sqrt{\det \boldsymbol{R}_x}\sqrt{\det \boldsymbol{R}_x} = \sqrt{\left(1-R_i^2\right)}\sqrt{\left(1-R_j^2\right)}\sqrt{\det\left(\boldsymbol{R}_x\right)_{ii}}\sqrt{\det\left(\boldsymbol{R}_x\right)_{jj}} \tag{C.12}$$

For $i \neq j$ the *ij*-partial correlation coefficient is denoted by $r_{ij;12\cdots p}$, and defined [28]

$$r_{ij;12\cdots p} = (-1)^{i+j+1}\frac{\det\left(\boldsymbol{R}_x\right)_{ij}}{\sqrt{\det\left(\boldsymbol{R}_x\right)_{ii}}\sqrt{\det\left(\boldsymbol{R}_x\right)_{jj}}}$$

whereby for every $i, j = 1, 2, \cdots, p$, it is implied

$$\det\left(\boldsymbol{R}_x\right)_{ij} = (-1)^{i+j+1} r_{ij;12\cdots p}\sqrt{\det\left(\boldsymbol{R}_x\right)_{ii}}\sqrt{\det\left(\boldsymbol{R}_x\right)_{jj}} \tag{C.13}$$

Substituting Equation (C.12) and Equation (C.13) in Equation (C.10) arises:

$$\begin{aligned}
\operatorname{cov}\left(\beta_i, \beta_j\right) &= \frac{(-1)^{i+j}\sigma^2 r_{ij}}{(n-1)s_{x_i x_j}} \cdot \frac{\det\left(\boldsymbol{R}_x\right)_{ij}}{\det \boldsymbol{R}_x} \\
&= \frac{(-1)^{i+j}(-1)^{i+j+1}\sigma^2 r_{ij}}{(n-1)s_{x_i x_j}} \cdot \frac{\sqrt{\det\left(\boldsymbol{R}_x\right)_{ii}}\sqrt{\det\left(\boldsymbol{R}_x\right)_{jj}} r_{ij;12\cdots p}}{\sqrt{\left(1-R_i^2\right)}\sqrt{\left(1-R_j^2\right)}\sqrt{\det\left(\boldsymbol{R}_x\right)_{ii}}\sqrt{\det\left(\boldsymbol{R}_x\right)_{jj}}} \\
&= -\frac{\sigma^2 r_{ij}}{(n-1)s_{x_i x_j}} \cdot \frac{r_{ij;12\cdots p}}{\sqrt{\left(1-R_i^2\right)}\sqrt{\left(1-R_j^2\right)}}
\end{aligned}$$

Moreover, for $i = j$ combining Equations (C.9) and (C.11) the variance of β_i is derived as follows:

$$\operatorname{var}\left(\beta_i\right) = \frac{\sigma^2 \det\left(\boldsymbol{R}_x\right)_{ii}}{\sqrt{\sum \tilde{x}_i^2}^2 \det \boldsymbol{R}_x} = \frac{\sigma^2 \det\left(\boldsymbol{R}_x\right)_{ii}}{(n-1)s_{x_i}^2 \det \boldsymbol{R}_x} = \frac{\sigma^2}{(n-1)s_{x_i}^2\left(1-R_i^2\right)}$$

Appendix D

```
** Dataset concerning Diabetes in Pima Indians
** Several constraints were placed on the selection of these instances from a larger database.
** In particular, all patients here are females at least 21 years old of Pima Indian heritage.
** http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
** Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository
** [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California,
** School of Information and Computer Science.
** Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988).
** “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus”.
** In Proceedings of the Symposium on Computer Applications and Medical Care} (pp. 261--265).
** IEEE Computer Society Press.

** Plasma glucose concentration at 2 hours in an oral glucose tolerance test
** (this is the dependent variable here)
sum glucose
scalar n=r(N)
scalar y= r(mean)
scalar S11=r(Var)

** the independent variables
** Diastolic blood pressure (mm Hg)
** Triceps skin fold thickness (mm)
** 2-Hour serum insulin (mu U/ml)
** Body mass index (weight in kg/(height in m)^2)
** Age (years)
```

```
** obrain the correlation matrix of the predictors
corr  dbp thickness insulin bmi age
mat R22=r(C)
mat R22inv=invsym(R22)

** obtain the covariance matrix
corr  dbp thickness insulin bmi age,cov
mat S22=r(C)

** obtain the full covariance matrix
corr  glucose dbp thickness insulin bmi age,cov
mat S=r(C)
scalar col=colsof(S)
mat S12=S[1, 2..col]
mat S21=S[2..col, 1]
mat S22=S[2..col, 2..col]

**obtain the full correlation matrix
corr  glucose dbp thickness insulin bmi age
mat R=r(C)
mat Ryx=R[2..6, 1]
mat Rk=R[2..6, 2..6]
mat bs=invsym(Rk)*Ryx

** implementation of the Samsa, Hu and Root method
matrix A = J(1,5,0)
scalar k=1
foreach x in  dbp thickness insulin bmi age {
qui reg  glucose `x'
   mat bb=e(b)
   mat A[1, k]=bb[1,1]
   scalar k=k+1
}

local col2=col-1
matrix temp=vecdiag(S22)
matrix SS = J(1,5,0)

forvalues i=1(1) `col2' {
 mat SS[1,`i']=sqrt(temp[1,`i']  )
}

mat AS=hadamard(A,SS)
mat AAS= R22inv*AS'
matrix bs2 = J(5,1,0)

forvalues i=1(1) `col2' {
 mat bs2[`i',1]=AAS[`i',1]/SS[1,`i']
}

mat list bs2

** the standard method from multivariate analysis (the results are identical)
mat b=invsym(S22)*S21
mat list b

*calculation of sigma-squared
mat sigma2=S11-S12*b
** because this is estimated, we need to take it into account
scalar sigma2=(sigma2[1,1]*(n-1))/(n-6)

** calculation of R-squared for the independent variables
mat Sx=vecdiag(S22)
matrix R2 = J(1,5,0)

forvalues i=1(1) `col2' {
   mat R2[1,`i']=1-1/R22inv[`i' , `i' ]
}

scalar detR=det(R22)
```

```
** calculation of the Rkij matrix which contains the determinants of the R22 matrix removing each
** time a row and a column
matrix Rkij = J(5,5,0)
preserve
clear

forvalues i=1(1) `col2' {
  forvalues j=1(1) `col2' {
       qui svmat R22
       qui drop R22`i'
       qui drop in `j'
       qui mkmat R22*,mat(Rii`i'`j')

       mat Rkij[`i', `j']=det(Rii`i'`j')
       clear
  }
}
restore

** calculation of the Partial Correlation coefficients (stored in matrix Rp)
matrix Rp = J(5,5,0)

forvalues i=1(1) `col2' {
  forvalues j=1(1) `col2' {
       scalar ex=`i'+`j'+1
       mat Rp[`i', `j']=(-1)^(ex)*(Rkij[`i', `j']/(sqrt(Rkij[`i', `i'])*sqrt(Rkij[`j', `j'])))
  }
}

** calculation of the covariance matrix of the regression coefficients
** (stored finally in matrix Vb)
matrix seb = J(1,5,0)
forvalues i=1(1) `col2' {
  mat seb[1,`i']=sqrt(sigma2/((n-1)*S22[`i', `i']*(1-R2[1, `i'])))
}

mat vb=diag(seb)
mat Vb=-vb*Rp*vb
mat list Vb
reg glucose dbp thickness insulin bmi age
mat list e(V)
```