Statistical models in digital solutions in support of mental health in patients with CHF

Julia Volaufova*

1. Introduction

Congestive (chronic) heart failure (CHF) is a clinical syndrome characterized by the heart's inability to pump sufficient blood to meet the body demands. It often results in decrease of quality of life. It has been recognized as being associated with poor mental health outcomes, including depression, anxiety, and cognitive impairment. At the same time, mental health disorders may adversely affect disease management, while living with CHF may contribute to worsening of cognitive/psychiatric symptoms.

Clinicians, health care providers, and researchers need to investigate clinically and statistically meaningful associations between patient characteristics, such as measured bio-markers, physical characteristics, socioeconomic status, etc., predefined specific markers of CHF, and mental health conditions. The need is to monitor the progression of the disease and/or its connection to mental health and general well being of a patient. For that purpose, multiple linear regression models for continuous outcomes, and logistic regression models for dichotomous response, provide a robust framework to quantify such associations. *Prediction models* are developed, ideally as a team work including statisticians, medical professionals, computer scientists, etc., to study the relations in question. In prediction, the goal is to *estimate* the unobserved characteristic of the disease or a future observation of a marker based on a number of observed patients' characteristics (usually called predictors). Hence prediction involves primarily estimation in conjunction with probabilistic statements about the precision of the predicted value. Statistical models are most widely used in prediction (see Steyerberg (2009)).

Regression models are widely used in the prediction of outcomes related to CHF, due to their ability to model relationships between clinical variables and disease

*Professor Emerita, Biostatistics Program, School of Public Health, LSU Health-New Orleans, Louisiana, USA (jvolau@lsuhsc.edu).

outcomes. Regression models help estimate the probability or severity of congenital heart failure based on patient data such as: demographics (age, sex), comorbidities, clinical symptoms, diagnostic results (ECG, blood tests, etc.), genetic or prenatal screening data, socioeconomic status, medications, etc.

Common regression techniques include simple linear regression, which is useful in predicting continuous outcomes like ejection fraction or severity scores; multiple regression, which incorporates multiple predictors to improve accuracy and account for complex interactions, and logistic regression, which is used for binary classification (e.g., presence or absence of CHF, presence or absence of depression, etc.). Logistic regression models the probability of occurrence of an outcome based on input features.

In addition, often we have to take into consideration different sources of variability in order to better characterize the precision of the estimated relation between the outcome and predictors. For that, mixed linear, generalized linear or mixed nonlinear statistical models are used. In case of high-dimensional data, regularized regression (Lasso, ridge) techniques might be applied, which might prevent overfitting of a model.

Here we concentrate on regression models, estimation of parameters in those models, providing confidence intervals of the mean of the response as well as prediction intervals for future observations.

Throughout this report, vectors (matrices) will be denoted by bold letters (bold upper cases), $\mathcal{C}(\cdot)$ denotes the column vector space, and for any matrix $\boldsymbol{A}$. The transpose of a matrix (or a vector) is denoted by a prime, e.g.,as $\boldsymbol{A}'$. $P_{\boldsymbol{A}}$ denotes the orthogonal projection onto the column space $\mathcal{C}(\boldsymbol{A})$ or, without ambiguity the orthogonal projection matrix onto $\mathcal{C}(\boldsymbol{A})$. $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. $\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for distributed as a p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$. Random variables and random vectors are denoted by capital letters, their observed values by lower case letters. For a random vector (variable): $E[\cdot]$ stands for its expectation, $\mathrm{var}[\cdot]$ for its variance or for its dispersion matrix (variance) matrix, and $\mathrm{cov}[\cdot, \cdot]$ for the covariance matrix between two vectors (random variables). Other notation will be defined in the subsequent text as needed.

2. Linear regression

2.1 Estimation and prediction

In medical applications great deal of attention is given to continuous response, such as blood pressure, heart function score (HFS), etc. Also, various measures of quality of life or mental status are recorded on various scales, such as Beck de-

pression inventory, which strictly speaking provides ordinal response but without loss of efficiency can be treated as continuous response. Linear regression models are widely used in the analysis of those types of responses.

Let the objective be prediction of a single continuous outcome based on a single predictor, for example age. As an example, consider a study with n patients with observed HFS and age on each patient. Let the observed HFS on the i-th patient be denoted by y_i and the recorded age by x_i. The simple linear statistical model for the linear relationship between HFS and age then takes the form

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where β_0 and β_1 are the unknown regression parameters, ϵ_i is the error term with distributional assumptions given by the first two moments, expectation $E[\epsilon_i] = 0$ and variance $\text{var}[\epsilon_i] = \sigma^2$, for all $i = 1, 2, \ldots, n$. The focus is on estimation of β_0, β_1, and σ^2 and based on their estimators and their statistical properties, further prediction of the outcome may be provided. We differentiate between *estimator* as a random variable (or random vector) and *estimate*, its calculated value based on given data.

More generally, let multiple predictors, characteristics recorded on each patient, be available in order to more precisely develop a relation to the mean of a single outcome variable. Inputs might include age, blood pressure, cholesterol, ECG findings, etc., or functions of predictors, for example the square of age. In addition, certain categorical predictors may be included as predictors represented by so called dummy (or indicator) variables. Then the model, which captures the combined effect of several risk factors, is referred to as multiple regression model and takes the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ observable random vector containing observations of the continuous response on all n patients; $\boldsymbol{X}$ is an $n \times k$ known matrix of covariates, where each row, say $\boldsymbol{X}_{i.}$ corresponds to the k predictors observed on the i-th patient. The model for the mean of Y_i may be expressed as

$$E[Y_i] = \sum_{j=1}^{k} x_{ij}\beta_j = \boldsymbol{X}_{i.}\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n, \tag{2.3}$$

where x_{ij} denotes the $\{i, j\}$th entry of $\boldsymbol{X}$ and $\boldsymbol{X}_{i.}$ the ith row of $\boldsymbol{X}$, $\boldsymbol{\beta}$ is the vector of unknown mean parameters. The relation (2.3) means that the mean vector $E[\boldsymbol{Y}] \in \mathcal{C}(\boldsymbol{X})$. We assume that $E[\boldsymbol{\epsilon}] = \boldsymbol{0}$ and that observations on different subjects are statistically independent, which is expressed by the covariance matrix $\text{var}[\boldsymbol{Y}] = \sigma^2\boldsymbol{I}$ being proportional to an identity matrix. The vector $\boldsymbol{\beta} \in \mathcal{R}^k$ and the constant $\sigma^2 > 0$ are the unknown model parameters. Based on collected data,

we get the estimator of $E[\boldsymbol{Y}]$, $\widehat{\boldsymbol{Y}} = P_{\boldsymbol{X}}\boldsymbol{Y}$, from where the estimator of $\boldsymbol{\beta}$ is any k-vector $\widehat{\boldsymbol{\beta}}$, such that $P_{\boldsymbol{X}}\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. (See, e.g., LaMotte (2025).) The matrix $P_{\boldsymbol{X}}$ in this context is often referred to as the *hat matrix*.

The estimator of σ^2 is given by the expression

$$(n - \operatorname{tr}(P_{\boldsymbol{X}}))\widehat{\sigma^2} = \boldsymbol{Y}'(I - P_{\boldsymbol{X}})\boldsymbol{Y}. \tag{2.4}$$

Notice that the orthogonal projection matrix $P_{\boldsymbol{X}}$ is unique and may be obtained, e.g., as $P_{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{X}^{+}$, where $\cdot^{+}$ denotes the Moore-Penrose inverse of a matrix (see, e.g., Rao & Mitra (1971)). In most applications the columns of the matrix $\boldsymbol{X}$ are linearly independent and in such case we get $\boldsymbol{X}^{+} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, from where

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},\ E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta},\ \text{and}\ \operatorname{var}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2.5}$$

The estimators $\widehat{\boldsymbol{Y}}$ and $\widehat{\sigma^2}$ under our assumptions have optimal statistical properties.

It is essential to provide characterization of variability of an estimator $\widehat{\boldsymbol{Y}}$. The estimated mean $\widehat{\boldsymbol{Y}}$ is a linear function of the random vector $\boldsymbol{Y}$, hence its covariance matrix is $\operatorname{var}[\widehat{\boldsymbol{Y}}] = \sigma^2 P_{\boldsymbol{X}}$. The estimated covariance matrix is then obtained by plugging in the estimator of σ^2 from (2.4) into $\operatorname{var}[\widehat{\boldsymbol{Y}}]$ and then takes the form

$$\widehat{\operatorname{var}}[\widehat{\boldsymbol{Y}}] = \widehat{\sigma^2} P_{\boldsymbol{X}}. \tag{2.6}$$

It is straightforward to see that for the i-th subject the estimator of expectation of the response and its estimated variance are

$$\widehat{Y_i} = e_i' P_{\boldsymbol{X}}\boldsymbol{Y} \quad \text{and}\ \widehat{\operatorname{var}}[\widehat{Y_i}] = \widehat{\sigma^2} e_i' P_{\boldsymbol{X}} e_i,$$

where e_i is the i-th column of the $n \times n$ identity matrix. $e_i' P_{\boldsymbol{X}} e_i$ is the i-th diagonal entry of the hat matrix and may be denoted by h_{ii}.

Assume now that the error vector follows normal distribution with expectation and covariance matrix defined above, i.e., $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. If one is interested in testing a linear hypothesis about the expectation of the i-th subject's response, say, $H_0 : E[Y_i] = \mu_i$ against the alternative $H_a : E[Y_i] \neq \mu_i$, the test statistic takes the form

$$T_i(\boldsymbol{Y}) = \frac{e_i' P_{\boldsymbol{X}}\boldsymbol{Y} - \mu_i}{\sqrt{\widehat{\sigma^2} h_{ii}}}, \tag{2.7}$$

and under the null hypothesis it follows central t-distribution with $n - \operatorname{tr}(\boldsymbol{X})$ degrees of freedom, $t_{(n-\operatorname{tr}(\boldsymbol{X}))}$. Denote the observed value of $T_i(\boldsymbol{Y})$ by $T_i(\boldsymbol{y})$. The hypothesis H_0 is rejected if the calculated p-value, assuming H_0 is true, i.e.,

$p = P_0\left(|T_i(\boldsymbol{Y})| > |T_i(\boldsymbol{y})|\right)$ is less than the predefined level of significance, usually denoted by α (α may be 0.01, 0.05, 0.1, etc.). Let the significance level be set to a given α. Denote by $t_{(n-\mathrm{tr}(\boldsymbol{X}))}(\alpha/2)$ the upper $\alpha/2$ quantile of central $t_{(n-\mathrm{tr}(\boldsymbol{X}))}$ distribution. The hypothesis H_0 is rejected at level α if the observed vector $\boldsymbol{y}$ belongs to *rejection region* for H_0:

$$\mathcal{R}_i = \{\boldsymbol{y} \in \mathcal{R}^n : |T_i(\boldsymbol{y})| > t_{(n-\mathrm{tr}(\boldsymbol{X}))}(\alpha/2)\}. \tag{2.8}$$

From (2.7) and (2.8), we get the *acceptance region* in $\mathcal{R}^n$,

$$\mathcal{A}(\mu_i) = \{\boldsymbol{y} \in \mathcal{R}^n : \left|\frac{\widehat{y_i} - \mu_i}{\sqrt{\widehat{\sigma^2} h_{ii}}}\right| \leq t_{(n-\mathrm{tr}(\boldsymbol{X}))}(\alpha/2)\},$$

which, if the hypothesis H_0 is true, has probability $1-\alpha$, i.e.,

$$P_{\mu_i}(\mathcal{A}(\mu_i) = 1 - \alpha \tag{2.9}$$

and can be expressed also as

$$\mathcal{A}(\mu_i) = \left\{\boldsymbol{y} \in \mathcal{R}^n : \widehat{y_i} - t_{(n-\mathrm{tr}(\boldsymbol{X}))}(\alpha/2)\sqrt{\widehat{\sigma^2} h_{ii}} \leq \mu_i \leq \widehat{y_i} + t_{(n-\mathrm{tr}(\boldsymbol{X}))}(\alpha/2)\sqrt{\widehat{\sigma^2} h_{ii}}\right\}.$$

The probability in (2.9) holds true for every hypothesized μ_i. From that we immediately get the $(1-\alpha)$-**confidence interval** for the mean of the response for the i-th patient. It is the set of all values $\mu_i \in \mathcal{R}$, for which the hypothesis $H_0 : E(Y_i) = \mu_i$ is not rejected on level α, and is given as

$$\mathcal{C}_i(\boldsymbol{y}) = \left\{\mu_i \in \mathcal{R} : \widehat{y_i} - t_{(n-\mathrm{tr}(\boldsymbol{X}))}(\alpha/2)\sqrt{\widehat{\sigma^2} h_{ii}} \leq \mu_i \leq \widehat{y_i} + t_{(n-\mathrm{tr}(\boldsymbol{X}))}(\alpha/2)\sqrt{\widehat{\sigma^2} h_{ii}}\right\}.$$

Consider a setting where we have, as above, n independently sampled patients with recorded responses arranged in a vector $\boldsymbol{y}$ and the set of measured predictors arranged into a matrix $\boldsymbol{X}$. The relation between the mean of each Y_i and the set of predictors $x_{i,j},\ j = 1, 2, \ldots, k$ follows model (2.3). Assume that we have a new patient and that we are supposed to predict the value of this patient's unobserved response, (as an example we still may use HFS) based on a new set of recorded predictors, say $\boldsymbol{x}_* \in \mathcal{R}^k$. Denote the unobserved response variable by Y_0. We may use here a two-sample test set-up to derive the $(1-\alpha)$-**level prediction interval** for y_0.

Let $E[Y_0] = \mu_0$ and $\mathrm{var}[Y_0] = \sigma^2 (= \mathrm{var}[Y_i]$ for each $i)$. Denote by $\mu_* = \boldsymbol{x}_*'\boldsymbol{\beta}$. If $E[Y_0]$ followed the model (2.3) then μ_0 would be equal to μ_*. Set up a hypothesis $H_0 : \mu_0 = \mu_*$ vs. $H_a : \mu_0 \neq \mu_*$. The test statistic to test H_0 takes the form

$$T(\boldsymbol{Y}, Y_0) = \frac{|\widehat{\mu}_0 - \widehat{\mu}_*|}{\sqrt{\widehat{\mathrm{var}}[\widehat{\mu}_0 - \widehat{\mu}_*]}}.$$

It is straightforward to realize that $\widehat{\mu}_0 = Y_0$, $\widehat{\mu}_* = \boldsymbol{x}_*'\widehat{\boldsymbol{\beta}}$, and $\mathrm{var}[\widehat{\mu}_0 - \widehat{\mu}_*] = \sigma^2 + \mathrm{var}[\boldsymbol{x}_*'\widehat{\boldsymbol{\beta}}] = \sigma^2\left(1 + \boldsymbol{x}_*'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_*\right)$. Because $\widehat{\mu}_0 - Y_0 = 0$, the estimate $\widehat{\sigma^2}$ is given by (2.4). Following the same line of development as for the confidence interval for the ith mean, we obtain the $(1-\alpha)$-level prediction interval for y_0. These are all y_0 in the range

$$\boldsymbol{x}_*'\widehat{\boldsymbol{\beta}} \pm t_{(n-\mathrm{tr}(\boldsymbol{X}))}(\alpha/2)\sqrt{\widehat{\sigma^2}\left(1 + \boldsymbol{x}_*'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_*\right)}. \tag{2.10}$$

For more details see LaMotte and Volaufova (1999). The predicted value of y_0 is $\boldsymbol{x}_*'\widehat{\boldsymbol{\beta}}$. In every application, in addition to the predicted value, it is essential to provide also the corresponding $(1-\alpha)$-level prediction interval.

2.2 Variable selection

Up to this point, we assumed that the set of recorded predictors is given, i.e., that the statistical model is fully specified by the model equation and distributional properties, up to the unknown model parameters. In practice, however, it is not often the case. There is usually a wide collection of patients' characteristics that may be considered to be included in the model.

As the first step, if unknown from historic reasons, the functional relationship between the outcome and potential continuous predictor is usually established. Visual methods may be used for that purpose. As a result, in addition to additive (linear) terms, polynomials, exponential or trigonometric functions of a predictor, etc., or products of some predictors, may be considered in the model. Penalized B-splines provide a very general form-free approach for modeling effects of continuous predictors. For categorical predictors, such as economic status, it is important to establish the appropriate coding in order to avoid linear dependencies between the columns of the matrix $\boldsymbol{X}$.

In order to have a valid prediction model, one has to be aware of issues related to too many or too few predictors. While the full model including all available predictors seems to be reasonable, it might lead to reduced interpretability and overfitting. There does not exist a unique recipe for building a model. An important part of the process is medical validation of the relation between the outcome and potential predictor, i.e., a subject-matter knowledge. Particularly, patients' characteristics that have been established to have medically essential meaning in relation to the outcome variable, say in correct diagnostics, have to be *forced* to be included in every step of the model creation. In every step, the created model has to be evaluated. One of the widely used evaluation criteria, included in every statistical textbook, is the the coefficient of determination, R^2. It is defined as $R^2 = \left(1 - \dfrac{\boldsymbol{Y}'\left(\boldsymbol{I} - P_{\boldsymbol{X}}\right)\boldsymbol{Y}}{\boldsymbol{Y}'\left(\boldsymbol{I} - P_{\mathbf{1}}\right)\boldsymbol{Y}}\right)$, where $P_{\mathbf{1}}$ is the projection matrix onto the space generated by the vector of ones, $\mathbf{1}$. However, adding

another predictor – *any* predictor – to the model cannot decrease R^2. In multiple regression setting, there are several statistical methods and evaluation criteria that can be used. Each has its pros and cons.

Step-wise procedures examine effects of individual predictors on the model one at a time. In forward selection, one additional predictor is included in the model at each successive step. In backward elimination, one is removed at each step. Stepwise methods combine both approaches. For example, at each step, predictors in the model are considered for removal, and predictors not in the model are considered for inclusion. These methods have been around for many decades, going back at least to the 1940s. The decision to include or exclude the predictor under consideration is based on some criterion, most often considering the p-value that results from testing the regression parameter corresponding to that predictor against 0. The threshold for the p-value for inclusion or exclusion of the predictor is chosen subjectively and depends on a given setting. However, the step-wise procedures cannot guarantee optimality of the resulting model, which may result, e.g., in poor prediction performance of the model. For some details on the topic and illustrations see Peixoto & LaMotte (1989). In spite of some disadvantages, step-wise procedures are widely used and are implemented in most statistical software packages. In addition to the above mentioned p-value, criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) may be used. In clinical prediction models, AIC is somewhat preferred, see, e.g., Steyerberg (2009).

Another procedure that investigates models and sub-models is a so called *all possible subsets method*. If the number of all potential predictors is k, there are 2^k all possible subsets (models), to investigate, which with large k is not a feasible task. Instead, an algorithm created by Furnival and Wilson (1974) is incorporated into most statistical software packages. The algorithm identifies the best model among all of the same size. It is based on an idea that the *error sum of squares* defined for a given model as $SSE_p = \boldsymbol{Y}' \left(\boldsymbol{I} - P_{\boldsymbol{X}_p}\right) \boldsymbol{Y}$ is a monotone set function (here $\boldsymbol{X}_p$ is without loss of generality considered the model matrix for a given subset of p predictors). By adding variables into the model, SSE_p cannot increase, it usually decreases. The process, by comparison of already examined subsets, eliminates a large number of potential subsets of predictors from consideration. The question is then to identify among the best-of-a-given size models the most parsimonious one. Various model evaluation criteria are used for that purpose. In addition to AIC and BIC, Mallows' C_p is used, which in essence is similar to AIC. In the context of multiple regression, AIC for a model with p predictors, including the intercept, takes the form

$$AIC_p = n \ln\left(\frac{SSE_p}{n}\right) + p, \tag{2.11}$$

Lower values of AIC indicate better model. Mallows' C_p is here defined as

$$C_p = \frac{SEE_p}{\widehat{\sigma^2}} - (n - 2p). \tag{2.12}$$

Here $\widehat{\sigma^2}$ is the estimate of σ^2 based on the full model, containing all predictors. A "good" model based on Mallows' C_p criterion is such, for which $C_p \approx p$. It is worth mentioning that under our assumption of multiple regression and normal distribution of the error, and because all submodels are nested within the full model, minimization of AIC_p will result in the same model, for which $C_p \approx p$.

The selection methods mentioned above are widely used in medical practice. For large data sets and a large pool of potential predictors, methods based on log-likelihood functions can be used, such as LASSO, which estimates the model parameters and simultaneously (by shrinking) eliminates predictors from the model. For more details see Tibshirani (1996) and more recently Freijero-Gonzáles et al. (2022).

Overall, we may conclude this section by stating that it is very common to carry out some type of model building and then in the phase of inference (estimation, testing, and prediction), to pretend that the chosen model was the model that we started with. In such cases the process of inference is sequential, i.e., the same data are used to create a model and then to proceed with inference and hence the results are usually heavily biased toward "significant" outcomes. The preferred approach should be to create a model based on historical knowledge and sound medical reasoning regardless of recently observed data. For more detailed discussion see Steyerberg (2009).

3. Binary response

3.1 Estimation and prediction

When the presence or absence of CHF or of a mental health condition is investigated, the recorded outcome y_i for each patient is modeled by a dichotomous random variable, which takes only two values, say $y_i = 1$ for presence and $y_i = 0$ for absence of a condition. Denote by $\boldsymbol{x}_i$ the column vector of characteristics for the ith patient, covariates, $i = 1, 2, \ldots, n$, including the intercept. Let the conditional probability of "presence", given $\boldsymbol{x}_i$, be denoted by $P(Y = 1|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$ and from that $P(Y = 0|\boldsymbol{x}_i) = 1 - \pi(\boldsymbol{x}_i)$. In this setting, *logistic regression* is the method of choice for estimating the probability that a patient with certain characteristics exhibits presence of, say CHF, or, presence of a mental health disorder, etc. Logistic regression is a special, fairly simple, case of a *generalized linear model*, see, e.g., Hosmer, Lemeshow, and Sturdivant (2013) or Christensen (1997). Crucial in the study of logistic model is the *logit* transformation of $\pi(\boldsymbol{x}_i)$,

the logarithm of odds of the outcome, and assumption that the log-odds of the outcome is a linear function of predictors, presented by the model

$$logit\left(\pi(\boldsymbol{x}_i)\right) = \ln\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n, \tag{3.13}$$

where $\boldsymbol{\beta}$ is the vector of unknown model parameters.

Each coefficient β_j is interpreted as change in log-odds of having the feature under study, per unit increase in corresponding x_{ij}, keeping all other characteristics constant.

The response variables on different sampling units (patients) are assumed to be independent. Using Bernoulli distribution of the response and the log-likelihood function for $\boldsymbol{\beta}$, the estimation method is the maximum likelihood method. The estimate (MLE) of $\boldsymbol{\beta}$ is obtained by maximizing the log-likelihood function, using iterative process, which, at an appropriate stopping point, results in $\hat{\boldsymbol{\beta}}$. It is essential to note that the iterative equations are identical to normal equations in multiple regression for continuous response and under normal distribution. For details, see, e.g., Christensen (1997).

Assume now that we have an individual with observed characteristics given by a vector $\boldsymbol{x}_0$. If the outcome, say, y_0 were continuous, we would be interested in predicting the value y_0 using data and parameter estimates obtained from the set of fully observed n patients, as we presented in the multiple regression setting. In logistic regression we do not predict the outcome (1 or 0), we predict the probability $\pi(\boldsymbol{x}_0)$ of presence of the outcome event, assuming that the patient belongs to the same population as the others. From (3.13), the **predicted probability** takes the form

$$\hat{\pi}(\boldsymbol{x}_0) = \frac{1}{1+\exp(-\mathrm{x}_0'\hat{\boldsymbol{\beta}})}. \tag{3.14}$$

In addition to the predicted probability at $\boldsymbol{x}_0$ we have to provide the characterization of precision of the estimator. In order to express the variance and confidence interval for $\hat{\pi}(\boldsymbol{x}_0)$, we again proceed via the iterative process applying maximum likelihood estimation. Without going into technical details, denote, as in case of multiple regression, the $n \times k$- matrix containing all patients' characteristics by $\boldsymbol{X}$. Denote by $\widehat{\boldsymbol{D}}_0$ the $n \times n$ diagonal matrix with $\hat{\pi}(\boldsymbol{x}_i)(1-\hat{\pi}\left(\boldsymbol{x}_i\right))$ entries on the diagonal. From the Fisher information matrix we get the estimate of the large-sample covariance matrix for $\hat{\boldsymbol{\beta}}$, $\widehat{\mathrm{var}}[\hat{\boldsymbol{\beta}}] = \left(\boldsymbol{X}'\widehat{\boldsymbol{D}}_0\boldsymbol{X}\right)^{-1}$. Applying the Δ-method we get from that the approximate estimated variance of $\hat{\pi}(\boldsymbol{x}_0)$ as

$$\widehat{\mathrm{var}}\left[\hat{\pi}(\boldsymbol{x}_0)\right] \approx \hat{\pi}(\boldsymbol{x}_0)^2(1-\hat{\pi}(\boldsymbol{x}_0))^2\boldsymbol{x}_0'\widehat{\mathrm{var}}[\hat{\boldsymbol{\beta}}]\boldsymbol{x}_0. \tag{3.15}$$

There are multiple ways to test a hypothesis about $\pi(\boldsymbol{x}_0)$, say, $H_0 : \pi(\boldsymbol{x}_0) = p_*$. Our goal is to obtain the confidence interval for $\pi(\boldsymbol{x}_0)$. For its derivation, the

large-sample Wald test can be used. The resulting $(1-\alpha)$-confidence interval is then the set of all probabilities p_*, for which the hypothesis H_0 would not be rejected on a significance level α and is given by the range

$$\hat{\pi}(\boldsymbol{x}_0) \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}\,[\hat{\pi}(\boldsymbol{x}_0)]}, \tag{3.16}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of standard normal distribution. We emphasize that the interval (3.16) is a confidence interval, not a prediction interval for $\pi(\boldsymbol{x}_0)$. For further details see also Hosmer, Lemeshow, and Sturdivant (2013).

3.2 Variable selection

For model building, the same ideas as in multiple regression hold for logistic regression. One, in both cases, should keep in mind interpretability of predictors. The regression nature of the model dictates that the selection methods are basically the same as in multiple regression: step-wise methods, forward selection, backwards elimination, combination of the two, and also all-possible-subsets selection.

In step-wise procedures the p-values for inclusion (or exclusion) of variables (or sets of variables, in case of multi-categorical predictors) can be calculated from likelihood-ratio test or score test. It is worth mentioning that several statistical software packages, such as SAS® software, have the Furnival-Wilson algorithm for selecting best-of-a-given-size model incorporated also for logistic regression. The score function serves the purpose of a monotone set-function, analogous to the R^2 (or SSE) in multiple regression. In the final step, the most frequently used criterion for model choice among best-of-a-given size models is the AIC.

4. Mixed linear models

In clinical research it is a rare situation when patients are observed cross-sectionally, i.e., only once. The potential dependencies between observed random variables on each sampling unit - patient, in such case, have to be accounted for. Such dependencies occur when patients are observed longitudinally, over time, or, in a more complex way, if they are observed in multiple clinics, regions, etc. Often the observed variables are in *clusters*. In modeling the responses in such situations, in addition to fixed effects parameters we include also random cluster effects.

4.1 Random coefficients models

Assume that patients are followed longitudinally, with repeated observations, say over time. Examples are ejection fraction, BNP levels, or, repeated administration of mental health measures, etc. In order to capture progression of the

disease, or progression of psychological outcomes based on the state of the disease, models that capture heterogeneity among individuals and dependencies of observations on the same patient are applied. Typical models that are used in such settings are *random coefficients models*. There are several equivalent names for these types of models: random growth curve models, multi-level models, hierarchical models, etc.. Random coefficients models are special cases of so called *mixed models*. In these models, the outcome is captured for each patient multiple times. Each subject has his/her own trajectory over time, determined by easily interpretable function of time (linear, quadratic, etc.), with usually fewer parameters than the number of time points. The model parameters then are considered to be random - their randomness follows from the sampling scheme. Here the sampling unit is the patient. The model coefficients are linked then by a linear relation to patient's characteristics recorded in a vector, say, $\boldsymbol{x}_i$.

Following the notation and developments in Vonesh and Chinchilli (1997), we can present a simple form of the random coefficients model defined in two stages:

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i, \quad i = 1, \ldots, n; \tag{4.17}$$

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{Y}_i$, for each subject indexed by i, is an n_i dimensional response vector, containing the observations taken at n_i time points. Each row of the known matrix $\boldsymbol{Z}_i$ contains the intercept and functions of the times at which the responses are taken, which can be a linear term, square of the time, etc.. Joint multivariate normal distribution of the errors $\boldsymbol{e}_i$ and $\boldsymbol{b}_i$ is assumed as follows,

$$\begin{pmatrix} \boldsymbol{e}_i \\ \boldsymbol{b}_i \end{pmatrix} \sim N_{n_i+t} \begin{pmatrix} \sigma^2\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0}' & \boldsymbol{\Psi} \end{pmatrix}, \; i = 1, 2, \ldots, n,$$

i.e., the errors in the two stages are mutually independent. The positive definite variance matrix $\boldsymbol{\Psi}$ and the scalar variance $\sigma^2 > 0$ are unknown. One possible form of the known matrix $\boldsymbol{A}_i$ (but not limited to) can be presented as $\boldsymbol{A}_i = \boldsymbol{I}_t \otimes \boldsymbol{x}_i'$. The symbol $\otimes$ denotes the Kronecker product of matrices (vectors).

In such case, the expectation of each entry of the vector $\boldsymbol{\beta}_i$ is the linear combination of the entries of the q-vector of patients' characteristics, $\boldsymbol{x}_i$, with an unknown set of coefficients, population parameters. The entire population vector parameter $\boldsymbol{\beta}$ is tq dimensional. Assume that $\boldsymbol{Z}_i$ for all i are $n_i \times t$ known full column rank matrices. (For deeper discussion and more details about the model (4.17) see, e.g., Vonesh and Chinchilli (1997).) Realize that the marginal model takes the form

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad \text{var}[\boldsymbol{Y}_i] = \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i' + \sigma^2\boldsymbol{I}_{n_i}, \; i = 1, 2, \ldots, n. \tag{4.18}$$

Stacking the vectors $\boldsymbol{Y}_i$ one below the other, we get the column vector denoted by $\boldsymbol{Y}$. Denote $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{A}_i$. Stacking the matrices $\boldsymbol{X}_i$ one below the other results

in a matrix denoted by $\boldsymbol{X}$. The vector $\boldsymbol{e}$ is created in the same way from all $\boldsymbol{e}_i$ placed one below the other. Here we get the special form of *linear* mixed model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}, \tag{4.19}$$

with $E[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$, $\mathrm{var}[\boldsymbol{Y}] = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}_{\sum_{i=1}^n n_i}$, where $\boldsymbol{Z} = \mathrm{Diag}\{\boldsymbol{Z}_i\}$ and $\boldsymbol{D} = \boldsymbol{I} \otimes \boldsymbol{\Psi}$ are block diagonal matrices. We shall get into more details as we proceed.

Before any type of prediction in model (4.17) could be investigated, estimation and testing of population model parameters (and their linear functions) has to be carried out, as well as estimation of all second order parameters, and covariance matrices of all estimators have to be established. There is a vast literature on estimation and testing in linear mixed models. See, e.g., Vonesh and Chinchilli (1997), references therein, and many others.

Denote by $\boldsymbol{\theta}$ the vector containing all second order parameters, $\boldsymbol{\theta} = (\sigma^2, \mathrm{vech}(\boldsymbol{\Psi})')'$. (See, e.g., Vonesh and Chinchilli (1997).) The operation $\mathrm{vech}(\cdot)$ creates a column vector from a symmetric matrix placing the columns, starting from the diagonal elements, one below the other. Under the assumption of normality the established method of estimating, say $\boldsymbol{L}\boldsymbol{\beta}$, where $\boldsymbol{L}$ is a given matrix is by maximum likelihood (ML) method, which is an iterative process of simultaneous estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Another possible approach is by *generalized least squares method* (GLSE), by which we take the estimates of $\boldsymbol{\theta}$ obtained by *restricted maximum likelihood* method (REML), also an iterative process, and plug them into the least square solution for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ (obtained under a fixed $\boldsymbol{\theta}$). In both cases the large-sample properties of estimators are well established. For more details see, also, e.g., Harville (1976). In the special case of $\boldsymbol{A}_i = \boldsymbol{I} \otimes \boldsymbol{x}_i'$, yet another estimation method of σ^2 and $\boldsymbol{\psi}$ with good statistical properties is the method of moments (MM), by which in balanced models with observations taken at equal time points for all patients, the results coincides with the REML estimates.

Denote by $\hat{\boldsymbol{\beta}}$ the ML (or GLSE) estimator of $\boldsymbol{\beta}$ and by $\widehat{\boldsymbol{\theta}}$ the REML estimator of $\boldsymbol{\theta}$. In order to present the large sample variance $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) = \mathrm{var}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}]$ in our model (and subsequent results), we shall use, for convenience, additional notation (borrowed from Vonesh and Chinchilli (1997)). Let $\boldsymbol{\Sigma}_i(\boldsymbol{\theta}) = \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i' + \sigma^2\boldsymbol{I}$. Then, under a given $\boldsymbol{\theta}$, we get

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}|\boldsymbol{\theta}) = \left(\sum_{i=1}^n \boldsymbol{X}_i'\boldsymbol{\Sigma}_i(\boldsymbol{\theta})^{-1}\boldsymbol{X}_i\right)^{-1} = \left(\sum_{i=1}^n \boldsymbol{A}_i'(\boldsymbol{\psi} + \sigma^2(\boldsymbol{Z}_i\boldsymbol{Z}_i')^{-1})^{-1}\boldsymbol{A}_i\right)^{-1}.$$

(Recall that $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{A}_i$, for all i.) The large sample covariance matrix of $\hat{\boldsymbol{\theta}}$ can be obtained in two ways. From the restricted likelihood function one can obtain the Fisher information matrix (taking the negative of the expectation of

the Hessian matrix, the matrix of second derivatives of the log-likelihood), or as a good approximation, the negative of the Hessian. The large sample covariance matrix of $\hat{\boldsymbol{\theta}}$ is then the inverse of the Fisher information matrix (or of the Hessian). Let $\boldsymbol{E}_i(\boldsymbol{\theta}) = \dfrac{\partial \text{vec}(\boldsymbol{\Sigma}_i\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$ and $\boldsymbol{W}_i(\boldsymbol{\theta}) = 2\boldsymbol{\Sigma}_i(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_i(\boldsymbol{\theta})$. We plug in the estimates of $\boldsymbol{\theta}$ into the above expressions and then the large sample estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ can be presented as

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \boldsymbol{E}_i(\hat{\boldsymbol{\theta}})' \hat{\boldsymbol{W}}_i(\boldsymbol{\theta})^{-1} \boldsymbol{E}_i(\hat{\boldsymbol{\theta}}).$$

4.2 Prediction in model (4.19)

Assume that we are interested in predicting the unobserved value of the response Y. There might be two interesting scenarios.

Under Scenario 1, we would like to predict a value of Y_h for a subject, who was *not* included among the original n patients.

Under Scenario 2, we assume that the subject, say, the kth patient, was included among the original n patients. We are interested in predicting a value Y_{kh}, the observation, say, at a new time point indexed by h.

4.2.1 Scenario 1

Under this setting, we have to assume that the new patient follows the same model as the original n subjects. Denote by $\boldsymbol{A}_h$ the matrix that contains information about the characteristics of the new patient and by $\boldsymbol{z}_h$ the vector containing information about the time at which we want to predict the new patient's response. These assumptions lead to the model

$$Y_h = \boldsymbol{z}_h' \boldsymbol{A}_h \boldsymbol{\beta} + \boldsymbol{z}_h' \boldsymbol{b}_h + \boldsymbol{e}_h, \tag{4.20}$$

where the distributional assumptions for $\boldsymbol{b}_h$ and $\boldsymbol{e}_h$ are the same as in model (4. .18). We present here the linear unbiased predictor, which approximately minimizes the mean squared error in class of all unbiased predictors. From requirement of unbiasedness, it follows that the predicted Y_h takes the form

$$\hat{Y}_h = \boldsymbol{z}_h' \boldsymbol{A}_h \hat{\boldsymbol{\beta}},$$

and from (4.20) we get the large sample mean squared error, MSE as the variance of $\hat{Y}_h - Y_h$, i.e.,

$$E[\hat{Y}_h - Y_h]^2 = \text{var}[\hat{Y}_h - Y_h] = \boldsymbol{z}_h' \boldsymbol{A}_h \text{var}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] \boldsymbol{A}_h' \boldsymbol{z}_h + \boldsymbol{z}_h' \boldsymbol{\Psi} \boldsymbol{z}_h + \sigma^2.$$

The large sample estimated variance, $\widehat{\text{var}}[\hat{Y}_h - Y_h]$ is then obtained by plugging-in the REML estimators of $\boldsymbol{\theta}$ into the above expressions. Harville and Jeske (1992) argue that in small samples, the estimated variance underestimates the true variance, and suggested a correction term. In case that the number of patients is relatively large, and in order to simplify the approach, we advocate to use subsequently the large sample estimated variance without the correction term. Setting the significance level at α, the large sample $(1-\alpha)$-level prediction interval for y_h can be presented as all y_h in the range

$$\hat{Y}_h \pm t_{(n-q)}(\alpha/2)\sqrt{\boldsymbol{z}_h'\boldsymbol{A}_h\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\theta}})\boldsymbol{A}_h'\boldsymbol{z}_h + \boldsymbol{z}_h'\widehat{\boldsymbol{\Psi}}\boldsymbol{z}_h + \widehat{\sigma^2}}, \tag{4.21}$$

where $t_{(n-q)}(\alpha/2)$ denotes the upper $\alpha/2$ quantile of Student's central t-distribution with $n-q$ degrees of freedom.

4.2.2 Scenario 2.

If we wish to predict a new observation, say y_{kh}, for a patient whose data were included among the original n patient then we proceed as follows. The response vector $\boldsymbol{Y}_k$ for that k-th patient follows the model (4.17), i.e.,

$$\boldsymbol{Y}_k = \boldsymbol{Z}_k\boldsymbol{\beta}_k + \boldsymbol{e}_k,$$

$$\boldsymbol{\beta}_k = \boldsymbol{A}_k\boldsymbol{\beta} + \boldsymbol{b} + k.$$

Let $\boldsymbol{z}_{kh}$ be a vector containing information about the h-th time point at which we wish to predict the value y_{kh}. Denote by $\hat{\boldsymbol{b}}_k$ the best linear unbiased predictor (BLUP) of $\boldsymbol{b}_k$ at $\boldsymbol{\theta}$, which takes the form (see, e.g., Harville (1976)),

$$\hat{\boldsymbol{b}}_k = \boldsymbol{\Psi}\boldsymbol{Z}_k'\left(\boldsymbol{Z}_k\boldsymbol{\Psi}\boldsymbol{Z}_k' + \sigma^2\boldsymbol{I}\right)^{-1}\left(Y_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right).$$

The large sample variance at $\boldsymbol{\theta}$ of $\hat{\boldsymbol{b}}_k - \boldsymbol{b}_k$ is then

$$\text{var}[\hat{\boldsymbol{b}}_k - \boldsymbol{b}_k] = \boldsymbol{\Psi} - \boldsymbol{\Psi}\boldsymbol{Z}_k'\boldsymbol{\Sigma}_k(\boldsymbol{\theta})^{-1}\boldsymbol{Z}_k\boldsymbol{\Psi} + \boldsymbol{\Psi}\boldsymbol{Z}_k'\boldsymbol{X}_k\boldsymbol{V}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_k'\boldsymbol{\Sigma}_k(\boldsymbol{\theta})^{-1}\boldsymbol{Z}_k\boldsymbol{\Psi}.$$

Under scenario 2, the predicted y_{kh} is given by

$$\hat{y}_{kh} = \boldsymbol{z}_{kh}'\boldsymbol{A}_k\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}) + \boldsymbol{z}_{kh}'\hat{\boldsymbol{b}}_k.$$

The large sample variance of $\hat{y}_{kh} - y_{kh}$ is

$$\text{var}[\hat{y}_{kh} - y_{kh}] = \text{var}\left[\boldsymbol{x}_{kh}'(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{\beta}) + \boldsymbol{z}_{kh}'(\hat{\boldsymbol{b}}_k - \boldsymbol{b}_k) - \boldsymbol{e}_{kh}\right]. \tag{4.22}$$

Let

$$\boldsymbol{V}_k = \text{var}[\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{\beta}].$$

Denote the matrix of covariances by $\boldsymbol{C}_k$, then

$$\boldsymbol{C}_k = \text{cov}\left[(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{\beta}), (\hat{\boldsymbol{b}}_k - \boldsymbol{b}_k)'\right] = -\boldsymbol{V}(\hat{\boldsymbol{\beta}}|\boldsymbol{\theta})\boldsymbol{X}_k'\boldsymbol{\Sigma}_k(\boldsymbol{\theta})^{-1}\boldsymbol{Z}_k\boldsymbol{\Psi}.$$

Using the above expression in the evaluation of expression (4.22), we get

$$\text{var}[\hat{y}_{kh} - y_{kh}] = \boldsymbol{x}_{kh}'\boldsymbol{V}(\hat{\boldsymbol{\beta}}|\boldsymbol{\theta})\boldsymbol{x}_{kh} + \boldsymbol{z}_{kh}'\boldsymbol{V}_k\boldsymbol{z}_{kh} + \boldsymbol{x}_{kh}'\boldsymbol{C}_k\boldsymbol{z}_{kh} + \boldsymbol{z}_{kh}'\boldsymbol{C}_k'\boldsymbol{x}_{kh} + \sigma^2. \quad (4.23)$$

The approximation to the estimated large sample variance $\widehat{\text{var}}[\hat{y}_{kh} - y_{kh}]$, is then obtained by plugging in the estimates $\boldsymbol{\theta}$ parameters. The $(1-\alpha)$-level prediction interval for y_{kh} is given as the range of values belonging to

$$\hat{y}_{kh} \pm t_{(n-q)}(\alpha/2)\sqrt{\boldsymbol{x}_{kh}'\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\theta}})\boldsymbol{x}_{kh} + \boldsymbol{z}_{kh}'\hat{\boldsymbol{V}}_k\boldsymbol{z}_{kh} + \boldsymbol{x}_{kh}'\hat{\boldsymbol{C}}_k\boldsymbol{z}_{kh} + \boldsymbol{z}_{kh}'\hat{\boldsymbol{C}}_k'\boldsymbol{x}_{kh} + \widehat{\sigma^2}}.$$

5. Logistic regression with random effects

Consider a dichotomous response, such as presence or absence of CHF or mental condition. In many practical situations the response variables are not independent, similarly, as in the linear mixed model, but clustered. For example, if patients (subjects) are observed over time then the observations on a given subject constitute a cluster. If the subjects are observed in multiple clinics, then the observations are clustered within a clinic, etc.. In such cases random effects are introduced into the logistic regression, very similarly as in case of a continuous response. The logistic regression model with random effects is then a special case of a *generalized linear mixed model*.

Let, in general terms, the observations be grouped into clusters indexed by j, $j = 1, 2, \ldots, N$. Let for the jth cluster be n_j dichotomous observations of the response Y_{ij} and let the corresponding vector of covariates be $\boldsymbol{x}_{ij}$. In addition, consider a random effects vector $\boldsymbol{b}_j$ associated with the j-th cluster and a corresponding vector of covariates, $\boldsymbol{z}_{ij}$. Denote by $\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_j$. The conditional probability of "presence" of a condition (depression, other mental condition, CHF, etc.), is Bernoulli with probability $P(Y_{ij} = 1|\boldsymbol{b}_j)$,

$$P(Y_{ij} = 1|\boldsymbol{b}_j) = \frac{\exp(\eta_{\text{ij}})}{1 + \exp(\eta_{\text{ij}})}, \quad (5.24)$$

end hence the logistic model equation is

$$\log \frac{\pi(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_j)}{1 - \pi(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_j)} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_j.$$

Here we assume that $\boldsymbol{b}_j$ follows multivariate normal distribution, $N(\boldsymbol{0}, \Sigma)$, Σ an unknown positive definite matrix, often parametrized by fewer parameters, say

a vector $\boldsymbol{\theta}$. For estimation of all unknown parameters the marginal likelihood function is used, which in this case takes the form

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{j=1}^{J} \int \left[\prod_{i=1}^{n_j} f(y_{ij}|\boldsymbol{b}_j, \boldsymbol{\beta}) \right] \phi(\boldsymbol{b}_j|\boldsymbol{\theta}) d\boldsymbol{b}_j. \tag{5.25}$$

Here $f(y_{ij}|\boldsymbol{b}_j, \boldsymbol{\beta})$ denotes the probability function of Bernoulli distribution and $\phi(\boldsymbol{b}_j|\boldsymbol{\theta})$ denotes the probability density function of multivariate normal distribution. For maximization of (5.25) numerical integration methods have to be applied, these are implemented in statistical software packages, such as, for example, SAS.

We want to emphasize here that in our approach, the central goal in model (5.24), is the prediction of the cluster-specific probability, interpretable, for example, as a probability of a response observed at a future time for a given subject (patient, cluster). Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood estimate of $\boldsymbol{\beta}$ and let $\widehat{\boldsymbol{b}}_j$ be an optimal linear predictor or, preferably an empirical Bayes estimate of a random effect $\boldsymbol{b}_j$. The cluster-specific *predicted conditional probability* is given as

$$\widehat{\pi}(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}|\widehat{\boldsymbol{b}}_j) = \frac{\exp(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\widehat{\boldsymbol{b}}_j)}{1 + \exp(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\widehat{\boldsymbol{b}}_j)}.$$

The estimated variance of $\widehat{\pi}(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}|\widehat{\boldsymbol{b}}_j)$ will be derived next. For brevity, we shall use the notation $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_j$. Then the estimated variance of $\widehat{\eta}_{ij} = \boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\widehat{\boldsymbol{b}}_j$ takes the form

$$\hat{\text{var}}[\widehat{\eta}_{ij}] = \boldsymbol{x}'_{ij}\text{var}[\widehat{\boldsymbol{\beta}}]\boldsymbol{x}_{ij} + \boldsymbol{z}'_{ij}\text{var}[\widehat{\boldsymbol{b}}_j]\boldsymbol{z}_{ij} + 2\boldsymbol{x}'_{ij}\text{cov}[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}_j]\boldsymbol{z}_{ij}, \tag{5.26}$$

in which expression $\text{cov}[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}_j]$ is often set to $\mathbf{0}$. For the fixed effects estimator, the covariance matrix is obtain as an inverse of the observed Fisher information matrix (Hessian). For $\text{var}[\widehat{\boldsymbol{b}}_j]$, $\widehat{\boldsymbol{b}}_j$ is obtained as $\widehat{\boldsymbol{b}}_j = E[\boldsymbol{b}_j|\boldsymbol{y}_j, \widehat{\boldsymbol{\theta}}]$, then its posterior variance is $\text{var}[\boldsymbol{b}_j|\boldsymbol{y}_j, \widehat{\boldsymbol{\theta}}]$. Applying the Δ-method, we get

$$\text{var}[\widehat{\pi}(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}|\widehat{\boldsymbol{b}}_j)] = \widehat{\pi}(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}|\widehat{\boldsymbol{b}}_j)(1 - \widehat{\pi}(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}|\widehat{\boldsymbol{b}}_j))\sqrt{\hat{\text{var}}[\widehat{\eta}_{ij}]}.$$

In order to construct a confidence interval for $\pi(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}|\widehat{\boldsymbol{b}}_j)$, we could proceed creating a Wald-type confidence interval. However, in order to guarantee the probability to be between the bounds $[0, 1]$, another approach is applied. Start with a large sample $(1 - \alpha)100\%$ confidence interval for η_{ij}. The lower bound of that interval given as $L_\eta = \widehat{\eta}_{ij} - z_{\alpha/2}\sqrt{\hat{\text{var}}[\widehat{\eta}_{ij}]}$ and the upper bound takes the form $U_\eta = \widehat{\eta}_{ij} + z_{\alpha/2}\sqrt{\hat{\text{var}}[\widehat{\eta}_{ij}]}$. Here $z_{\alpha/2}$ denotes the upper $\alpha/2$ critical value of

standard normal distribution. In the next step, the inverse of logit transformation is applied resulting in the interval

$$\pi(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}|\boldsymbol{b}_j) \in \left(\frac{\exp(L_\eta)}{1+\exp(L_\eta)}, \frac{\exp(U_\eta)}{1+\exp(U_\eta)} \right).$$

For more details see also McCulloch and Searle (2001).

6. Concluding remarks

The developments shown in this report present an overview of relevant topics and demonstrate a coherent progression of statistical methodology applicable in analysis of relation of CHF and mental characteristics of a patient. We began with classical multiple regression model culminating in logistic regression model with random effects. At each stage, the essential statistical tasks - estimation of parameters as well as random effects, prediction of future observations, development of confidence and/or prediction intervals with touching on variable selection - remain central. Undoubtedly, the methods become more sophisticated as the models increase in complexity.

6.1 Estimation across models

From the standpoint of estimation, the multiple linear regression model is distinguished by the closed-form nature of the solutions. Under normality, the least squares estimator coincides with maximum likelihood estimator. This exactness is gradually lost as the model becomes more general. In logistic regression, estimation requires iterative methods, such as Newton-Raphson or Fisher scoring, with the observed or expected Fisher information as basis for the variance of the estimator.

In linear mixed model, such as the here presented random coefficient model, maximum likelihood and restricted maximum likelihood are applied and these methods do not result in closed-form expressions, except balanced factorial models. Estimation of variance components requires inversion of block matrices involving both fixed and random effects. The same principle remains, the information matrix is the basis for expressing the precision of estimators.

Finally, in logistic regression with random effects, the likelihood function involves integral over the distribution of the random effects. Closed forms of solutions are not available. Numerical approximations, such as Laplace expansions or Gauss-Hermite quadrature are employed. The inference relies on approximations to the likelihood and its derivatives. The unifying principle is straightforward: estimation in all models is grounded in the maximization of the likelihood function. Asymptotic approximations provide theoretical justification for inference.

6.2 Prediction and inference

Prediction represents another line unifying the models. In linear regression, prediction intervals for future responses follow directly from normality and the structure of the variance of prediction error.

In logistic regression, one predicts the probability of "presence" of a condition, connected to the covariates via the logit link function. Confidence intervals for such probabilities are constructed via the inverse link of a linear predictor, $\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}$. The delta method applied to the logit link yields approximate intervals, while profile-likelihood methods provide refinements.

In linear mixed models, prediction takes on two distinct forms. One may predict a future value of a response, which requires accounting for both fixed and random sources of variation, or one may predict the random effects themselves. The best linear unbiased predictor (BLUP) provides an estimate of random effects, shrinking subject-specific estimates toward the overall mean in a manner that balances between-cluster and within-cluster variation.

This concept naturally extends to logistic regression with random effects. The conditional probability of "success" here depends on both fixed covariates and subject- or cluster-specific random effects. Prediction here involves not only estimating fixed parameters (coefficients) but also integrating or conditioning on random effects. Confidence intervals for predicted probabilities are more approximate in this setting, as both the non-linearity of the link and the uncertainty in random effects must be accounted for. Nonetheless, the general philosophy persists: prediction must propagate all sources of uncertainty, ensuring that inferential statements about future observations reflect both fixed and random components of the model.

6.3 Variable selection and model building

Variable selection and model building are likewise recurring themes. In multiple regression, information criteria such as AIC, C_p, or BIC can be applied directly, with computational algorithms such as all-subset regression available to explore model space.

In logistic regression, similar criteria apply, though the absence of an exact analogue of C_p emphasizes reliance on likelihood-based approaches. Penalization methods, such as lasso or ridge regression, extend naturally to logistic models as well.

In linear mixed models and logistic regression models with random effects, model selection is much more delicate. Correlation structures and variance components complicate the effective sample size and the inclusion or exclusion of random effects substantially alters the interpretation of fixed effects. For example, treat-

ing a coefficient as random variable rather than fixed changes the dimension of parameter space and impacts the shrinkage of subject-specific estimates. Information criteria remain useful but must be interpreted with caution, as likelihoods in mixed models are evaluated with respect to both fixed and random parameters.

Thus, while the guiding principles of parsimony and interpretability persist, their application requires grater care in mixed-model framework. Model building in this context must be guided by scientific reasoning, knowledge of the data-generating process, and awareness of the consequences of treating effects as fixed or random.

6.4 Strength of the mixed-model framework

One of the principal strengths of the mixed-model approach is its ability to accommodate correlation and hierarchical structure of observations. Classical regression assumes independence of observations, an assumption that fails in repeated measures, cluster designs, or longitudinal studies. Mixed models provide a natural remedy by explicitly modeling variability across clusters or subjects.

Random coefficients models illustrate this strength clearly: by allowing the coefficients (e.g., slopes and intercepts) to vary across subjects. They capture individual-specific trajectories without requiring an unmanageably large number of parameters. This balance of flexibility and parsimony is achieved through shrinkage, whereby subject-specific estimates are pulled toward the overall regression function in proportion to the relative magnitudes of within-subject and between-subject variability.

In logistic regression with random effects, these advantages carry over to binary outcomes. The inclusion of random intercepts accounts for unobserved heterogeneity among clusters, while random slopes allow differential effects of covariates across groups. This framework leads to more valid inferences about fixed effects, as ignoring correlation often results in underestimated standard errors and overly optimistic conclusions.

6.5 Limitations and open directions

Despite their strength, mixed models, particularly in the logistic regression setting, are not without limitations. Estimation relies heavily on numerical approximations, which may introduce bias in small samples or with complex random effects structures. Prediction intervals for probabilities are less standardized than in linear models, with competing methods yielding slightly different results. Variable selection in mixed models remains an area of active research, with no universally accepted criterion.

Furthermore, computational burden grows rapidly with complexity of the random

structure and the size of the data set. While modern software implements efficient algorithms, convergence problems and sensitivity to starting values remain practical challenges. These issues remind us that while mixed models generalize regression in elegant ways, they demand both statistical and computational care in their application.

6.6 Closing synthesis

The development from multiple regression through logistic regression, linear mixed models, and logistic regression with random effects illustrates the power and unity of the likelihood-based approach to statistical modeling. Though the details of estimation, prediction, and variable selection evolve as the models become more sophisticated, the underlying philosophy remains consistent. Each model represents a step in addressing increasingly complex data structures: from independent continuous responses, to binary outcomes, to correlated or hierarchical response variables. The general message is that statistical models must reflect the complexity of the research question and of the data they are intended to explain, while maintaining the balance between interpretability, rigor, and feasibility.

References

Harville, D.A.(1976). Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. *The Annals of Statistics.* Vol.4, pp. 384–395.

Harville, D.A. and Jeske, D.R. (1992). Mean Squared Error of Estimation or Prediction Under General Linear Model. *Journal of the American Statistical Association.* Vol.87, pp. 724–731.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X. (2013). Applied Logistic Regression. Third Edition. Wiley.

Christensen, R. (1997). Log-linear Models and Logistic Regression. Second Edition. Springer.

Freijero-Gonzáles, L., Febrero-Bande, M., and Gonzáles-Manteiga, W. (2022). A Critical Review of Lasso and Its Derivatives for Variable Selection under Dependence among Covariates. *International Statistical Review.* Vol.90, pp. 118–145.

Furnival, G.M. and Wilson, R.W. (1974). Regression by Leaps and Bounds. *Technometrics.* Vol.16, pp. 403–408.

LaMotte, L.R. (2025). Foundations of Multiple Regression and Analysis of Variance. To appear. Chapman and Hall.

LaMotte, L.R. and Volaufova, J. (1999). Prediction Intervals via Consonance Intervals. *The Statistician.* Vol.48, pp. 419–424.

McCulloch, C.E.and Searle, S.R. (2001). Generalized, Linear, and Mixed Models. Wiley.

Peixoto, J.L. and LaMotte, L.R. (1989). Simultaneous Identification of Outliers and Predictors Using Variable Selection Techniques. *Journal of Statistical Planning and Inference.*Vol. 23, pp.327–343.

Rao, C.R. & Mitra, S.K. (1971). Generalized Inverse of Matrices and its Applications. John Wiley & Sons, Inc..

Steyerberg, E.W. (2009). Clinical Prediction Models. Springer.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* Vol.58, pp. 267–288.

Vonesh, E.F. and Chinchilli, V.M. (1997). Linear and Nonlinear Models for the Analysis of Repeated Measurements. Marcel Dekker, Inc.