Weighted-Average Least Squares for Negative Binomial Regression

Kevin Huynh[a,*]

[a] *Faculty of Business and Economics, University of Basel, Peter Merian-Weg 6, 4052 Basel, Switzerland*

Abstract

Model averaging methods have become an increasingly popular tool for improving predictions and dealing with model uncertainty, especially in Bayesian settings. Recently, frequentist model averaging methods such as information theoretic and least squares model averaging have emerged. This work focuses on the issue of covariate uncertainty where managing the computational resources is key: The model space grows exponentially with the number of covariates such that averaged models must often be approximated. Weighted-average least squares (WALS), first introduced for (generalized) linear models in the econometric literature, combines Bayesian and frequentist aspects and additionally employs a semiorthogonal transformation of the regressors to reduce the computational burden. This paper extends WALS for generalized linear models to the negative binomial (NB) regression model for overdispersed count data. A simulation experiment and an empirical application using data on doctor visits were conducted to compare the predictive power of WALS for NB regression to traditional estimators. The results show that WALS for NB improves on the maximum likelihood estimator in sparse situations and is competitive with lasso while being computationally more efficient.

Keywords: WALS, model averaging, negative binomial regression, count data
JEL Classification: C51, C25, C13, C11

*Corresponding author: Kevin Huynh, Faculty of Business and Economics, University of Basel, Peter Merian-Weg 6, 4052 Basel, Switzerland, E-Mail: kevin.huynh@unibas.ch

1 Introduction

In many empirical applications, model uncertainty emerges for a variety of reasons. For example, competing theories exist that can describe the data, or different assumptions are imposed on the data-generating process (DGP). The two most common approaches for dealing with model uncertainty are model selection and model averaging. In model selection, the user selects the best performing model according to an estimation criterion and then carries out inference based on the chosen model. This approach is problematic because the uncertainty in the initial model selection step is often ignored, which could lead to overly confident decisions and predictions (Steel, 2020). In contrast, model averaging accounts for model uncertainty by averaging over a set of candidate models, typically aiming at improving predictive accuracy (Ando and Li, 2014).

As datasets become larger, researchers commonly find themselves in high-dimensional settings with many potential covariates to model their response variable. Choosing appropriate regressors is particularly difficult in these situations because the number of candidate models grows exponentially with the number of regressors, i.e. for k regressors, 2^k different subsets exist that may be considered as candidates. For the same reason, managing the model space and computational resources is key to applying model averaging procedures in the presence of covariate uncertainty. Bayesian model averaging (BMA) provides two general approaches: 1. Markov chain Monte Carlo methods (MCMC) and 2. non-MCMC approximation methods, see e.g. Hoeting et al. (1999, p. 384 ff.) for an early overview. A common solution adopted in frequentist model averaging (FMA), e.g. in Zhang et al. (2016), is to prescreen for a viable set of models. In contrast, weighted-average least squares (WALS), first proposed by Magnus et al. (2010) for the linear regression model and then extended by De Luca et al. (2018) to generalized linear models (GLMs), omits a preselection of models by combining Bayesian and frequentist aspects and, especially, leveraging a semiorthogonal transformation of the regressors allowing for fast computation times. Earlier work by Heumann and Grenke (2010) generalizes WALS to logistic regression using a similar transformation as in De Luca et al. (2018).

Most of the literature, particularly in economics, has focused on model averaging for linear regression models. However, many interesting applications require nonlinear models, e.g. classification, count data modeling and survival analysis. The negative binomial (NB) distribution, especially of type 2 (NB2), is a popular distribution featuring overdispersion for count data regression, see e.g. Cameron and Trivedi (1986) and Cameron et al. (1988) for applications in health economics. Notably, the NB2 regression model is not a GLM when its dispersion parameter is estimated from the data. Deb and Trivedi (2002) extend it to hurdle and finite mixture models and Greene (2008) develops a more general form, called NBP, which encompasses the NB of type 1 and 2.

Despite its wide application, very limited literature exists on model averaging methods for the NB regression model that jointly estimate the regression coefficients and the dispersion parameter. One of the few open-source packages for model averaging is BMA by Raftery et al. (2020), which currently supports BMA for GLMs and survival models. Hence, it is only able to fit an NB2 with pre-specified dispersion parameter, which is a GLM.

In this paper, I extend WALS GLM by De Luca et al. (2018) to the NB2 regression model (WALS NB) to account for covariate uncertainty in the specification of the linear predictor. WALS is particularly well suited as it elegantly circumvents a preselection of models by transforming the regressors, allowing me to focus on the averaging procedure. Analogous to De Luca et al. (2018), I first derive the one-step maximum likelihood estimator based on a Taylor expansion of the NB2 log-likelihood function and then employ a transformation akin to the semiorthogonal transformation used in WALS GLM.

At the time of writing, the asymptotic distribution of the WALS estimator for GLMs is still an open research topic and its variance estimator has been a subject of debate. Recent work by De Luca et al. (2022) proposes a new estimator for the variance of WALS in the linear regression model instead of the Bayesian posterior variance that has traditionally been used. De Luca et al. (2023) further analyze the confidence and prediction intervals of WALS in the linear model and propose a new simulation-based method that corrects for bias in the WALS estimator. In contrast, this work focuses on the predictive power of model averaging and leaves the challenging issue of inference (after model averaging) for future research. Model averaging estimators typically improve the predictive accuracy compared to using a single model. For example, in an early application of BMA, Madigan and Raftery (1994) find that BMA achieves better logarithmic predictive score than any single model. Moreover, Min and Zellner (1993) show that the expected squared error loss of predictive mean forecasts is always minimized by BMA, if the data-generating model is included in the model space considered for averaging. In this paper, I compare the proposed WALS NB method to traditional maximum likelihood (ML) estimation of the NB2 regression model in a simulation experiment using the classical precision measure, root mean squared error (RMSE), and scoring rules (Gneiting and Raftery, 2007) as measures for the distributional fit. Finally, the method is also compared to the lasso estimator (Wang et al., 2016) in an empirical application on modeling doctor visits. Both the simulation experiment and the empirical application show that WALS NB improves on the ML estimator in sparse situations with few observations and many covariates. In the latter, its fit is competitive with lasso while being computationally more efficient.

2 Setup

The setup and derivation of WALS NB mostly follow the steps in De Luca et al. (2018) for WALS GLM. Assume that data $y_i, i = 1, 2, \ldots, n$, are conditionally independent given k-dimensional regressors x_i and follow an NB2 distribution with mean μ_i and dispersion parameter ρ, i.e. $y_i|x_i \sim \text{NB2}(\mu_i, \rho)$. As in the standard GLM setup, I model the mean using an inverse link function h on $\mu_i := \mu(\eta(\beta, x_i)) = h(\eta(\beta, x_i))$ with linear predictor $\eta_i := \eta(\beta, x_i) = x_i^\top \beta$ and regression coefficients β. The NB2 distribution has the probability mass function

$$f(y_i|\mu_i, \rho) = \frac{\Gamma(y_i + \rho)}{\Gamma(\rho)\Gamma(y_i + 1)} \frac{\mu_i^{y_i} \rho^\rho}{(\mu_i + \rho)^{y_i + \rho}}, \quad y_i \in \mathbb{N}_0,\ \rho > 0, \tag{2.1}$$

where Γ is the gamma function, and its conditional variance is given by

$$\sigma_i^2 := \mathsf{Var}(y_i|\mu_i, \rho) = \mu_i + \frac{\mu_i^2}{\rho}. \tag{2.2}$$

A distribution from the exponential family has the following density

$$f(y_i|\theta_i) = \exp(y_i\theta_i - b(\theta_i) + l(y_i)),$$

where b and l are known functions. Typical formulations as in e.g. Fahrmeir et al. (2013, p. 301) include a dispersion parameter which, without loss of generality, I set equal to one. Moreover, the following two identities hold for the mean and variance

$$\mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}, \qquad \sigma_i^2 = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}.$$

For WALS estimation, I rewrite the NB2 probability mass function into a similar form as the exponential family with a log-link on ρ by using

$$\theta_i := \theta\left(\mu_i, \rho(\alpha)\right) = \theta\left(h\left(\eta(\beta, x_i)\right), \rho(\alpha)\right) = \log\left(\frac{\mu_i}{\mu_i + \rho}\right) = \log(\mu_i) - \log(\mu_i + \rho),$$
$$\rho(\alpha) = \exp(\alpha).$$

Thus, the probability mass function becomes

$$f(y_i|\theta_i, \rho) = \exp\left(y_i\theta_i + \rho\log(1 - \exp(\theta_i)) + \log\Gamma(y_i + \rho) - \log\Gamma(\rho) - \log\Gamma(y_i + 1)\right),$$

where I dropped the dependence of θ on μ and ρ, and of ρ on α for notational brevity. From the last line, we can identify the following building blocks of the exponential family:

$$b(\theta_i, \rho) = -\rho\log(1 - \exp(\theta_i)),$$
$$l(y_i, \rho) = \log\Gamma(y_i + \rho) - \log\Gamma(\rho) - \log\Gamma(y_i + 1).$$

Thus, for fixed ρ, the NB2 is a member of the exponential family and leads to a GLM. However, ρ is estimated from the data in the WALS procedure and, hence, the underlying model is not a GLM anymore. Furthermore, I separate $l(y_i, \rho)$ into two terms using

$$a(y_i, \rho) := \log\Gamma(y_i + \rho) - \log\Gamma(\rho), \qquad d(y_i) := -\log\Gamma(y_i + 1),$$

so the NB2 probability mass function can be rewritten as

$$f(y_i|\theta_i, \rho) = \exp(y_i\theta_i - b(\theta_i, \rho) + l(y_i, \rho)) = \exp(y_i\theta_i - b(\theta_i, \rho) + a(y_i, \rho) + d(y_i)),$$

which will simplify the derivation of the WALS estimator later.

I allow for uncertainty in the specification of the linear predictor while assuming that the (conditional) probability mass function of y_i and the inverse link h are correctly specified. First, collect over all observations n the response y_i to an n-vector y and the regressors x_i

to an $n \times k$ matrix X that contains $x_i^\top$ as ith row. Then, partition the regressors into focus and auxiliary regressors $X = (X_1, X_2)$, where X_p is an $n \times k_p$ matrix with ith row equal to $x_{ip}^\top, p = 1, 2$, and $k_1 + k_2 = k$. Further, let $\beta = (\beta_1^\top, \beta_2^\top)^\top$ so the linear predictor can be expressed as $\eta_i = x_{i1}^\top \beta_1 + x_{i2}^\top \beta_2$. Stacking the linear predictors over all n observations then gives the vector $\eta(\beta) = X_1\beta_1 + X_2\beta_2$.

Consider averaging over models containing all focus regressors X_1 but arbitrary subsets of the k_2 auxiliary regressors in X_2, which leads to a total of 2^{k_2} possible models. The jth model is represented by the restriction $R_j^\top \beta_2 = 0$, where R_j denotes a $k_2 \times r_j$ matrix of rank $0 \leq r_j \leq k_2$, such that $R_j^\top = (I_{r_j}, 0)$ or column-permutations thereof. Thus, the matrix R_j specifies which auxiliary regressors are excluded from the jth model and its rank r_j denotes the number of excluded auxiliary regressors. Note that 0 represents a scalar, vector or matrix filled with zeroes of matching dimension unless otherwise stated. For example, 0 in $R_j^\top = (I_{r_j}, 0)$ is an $r_j \times (k_2 - r_j)$ matrix.

3 ML estimation

I start with the classical maximum likelihood estimator of the NB2 regression model. Under conditional independence, the (conditional) log-likelihood is

$$\begin{aligned} \ell(\beta, \alpha) &= \sum_{i=1}^{n} \log f\big(y_i | \theta(\mu_i(\beta), \rho(\alpha)), \rho(\alpha)\big) = \sum_{i=1}^{n} \left[y_i\theta_i - b(\theta_i, \rho) + a(y_i, \rho) + d(y_i)\right] \\ &= \text{constant} + \sum_{i=1}^{n} \left[y_i\theta_i - b_i + a_i\right], \end{aligned} \tag{3.1}$$

where $b_i := b(\theta_i, \rho)$ and $a_i := a(y_i, \rho)$. In the following, I will generally omit the dependence of θ, μ, η, ρ, b and a on their parameters to reduce clutter. Moreover, only the log-link is considered for the mean (and dispersion) parameter, i.e. $h(\eta_i) = \exp(\eta_i)$, but the general notation using h is retained in many places below to facilitate comparisons with WALS GLM by De Luca et al. (2018) and to allow easier extension of the method to other link functions in the future.

The score functions follow:

$$\begin{aligned} s_p(\beta, \alpha) &:= \frac{\partial \ell(\beta, \alpha)}{\partial \beta_p} = \sum_{i=1}^{n} \left[y_i \frac{\partial \theta_i}{\partial \eta_i} - \frac{\partial b_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \eta_i}\right] x_{ip} = \sum_{i=1}^{n} v_i[y_i - \mu_i]x_{ip}, \quad p = 1, 2, \\ s_\alpha(\beta, \alpha) &:= \frac{\partial \ell(\beta, \alpha)}{\partial \alpha} = \sum_{i=1}^{n} \left[y_i \frac{\partial \theta_i}{\partial \rho} - \frac{\partial b_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \rho} - \frac{\partial b_i}{\partial \rho} + \frac{\partial a_i}{\partial \rho}\right] \frac{\partial \rho}{\partial \alpha} = \sum_{i=1}^{n} \kappa_i \frac{\partial \rho}{\partial \alpha}, \end{aligned}$$

with

$$\begin{aligned} v_i &:= v(\eta_i, \rho) := \frac{\partial \theta_i}{\partial \eta_i}, \\ \kappa_i &:= \kappa(\eta_i, \rho, y_i) := y_i \frac{\partial \theta_i}{\partial \rho} - \frac{\partial b_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \rho} - \frac{\partial b_i}{\partial \rho} + \frac{\partial a_i}{\partial \rho} \\ &= -\frac{y_i - \mu_i}{\mu_i + \rho} + \log(\rho) - \log(\mu_i + \rho) + \text{di}(y_i + \rho) - \text{di}(\rho), \end{aligned}$$

where $\mathrm{di}(x) := \partial \log \Gamma(x)/\partial x$ is the digamma function. Furthermore, let $H(\beta, \alpha)$ be the negative Hessian of the log-likelihood, which is composed of several submatrices that are listed below. The first components are

$$\begin{aligned} H_{pq}(\beta,\alpha) &:= -\frac{\partial^2 \ell(\beta,\alpha)}{\partial\beta_p \partial\beta_q^\top} \\ &= -\sum_{i=1}^n \left[y_i \frac{\partial^2\theta_i}{\partial\eta_i^2} - \left(\frac{\partial^2 b_i}{\partial\theta_i^2} \left(\frac{\partial\theta_i}{\partial\eta_i} \right)^2 + \frac{\partial b_i}{\partial\theta_i} \frac{\partial^2\theta_i}{\partial\eta_i^2} \right) \right] x_{ip} x_{iq}^\top \\ &= \sum_{i=1}^n \left[v_i^2 \sigma_i^2 - \omega_i (y_i - \mu_i) \right] x_{ip} x_{iq}^\top = \sum_{i=1}^n \psi_i x_{ip} x_{iq}^\top, \quad p,q = 1,2, \end{aligned}$$

where

$$\omega_i := \omega(\eta_i, \rho) := \frac{\partial^2\theta_i}{\partial\eta_i^2}, \qquad \psi_i := \psi(\eta_i, \rho, y_i) := v_i^2\sigma_i^2 - \omega_i(y_i - \mu_i).$$

The next submatrices are defined as

$$\begin{aligned} H_{p\alpha}(\beta,\alpha) &:= -\frac{\partial^2 \ell(\beta,\alpha)}{\partial\beta_p \partial\alpha} \\ &= \sum_{i=1}^n \left[-y_i \frac{\partial^2\theta_i}{\partial\eta_i\partial\rho} + \left(\frac{\partial^2 b_i}{\partial\theta_i^2} \frac{\partial\theta_i}{\partial\rho} + \frac{\partial^2 b_i}{\partial\theta_i\partial\rho} \right) \frac{\partial\theta_i}{\partial\eta_i} + \frac{\partial b_i}{\partial\theta_i} \frac{\partial^2\theta_i}{\partial\eta_i\partial\rho} \right] x_{ip} \frac{\partial\rho}{\partial\alpha} \\ &= H_{\alpha p}(\beta,\alpha)^\top, \quad p = 1,2. \end{aligned}$$

They further simplify thanks to

$$\frac{\partial^2 b_i}{\partial\theta_i^2} \frac{\partial\theta_i}{\partial\rho} + \frac{\partial^2 b_i}{\partial\theta_i\partial\rho} = -\left(\frac{\mu_i^2}{\rho} + \mu_i \right) \frac{1}{\mu_i + \rho} + \frac{\mu_i}{\rho} = 0, \tag{3.2}$$

so $H_{p\alpha}(\beta, \alpha)$ may be rewritten as

$$H_{p\alpha}(\beta,\alpha) = -\sum_{i=1}^n c_i [y_i - \mu_i] x_{ip} \frac{\partial\rho}{\partial\alpha}, \quad p = 1,2,$$

using $c_i := c(\eta_i, \rho) = \partial^2\theta_i/\partial\eta_i\partial\rho$. Finally, the last part is

$$H_{\alpha\alpha}(\beta,\alpha) := -\frac{\partial^2\ell(\beta,\alpha)}{\partial\alpha^2} = -\sum_{i=1}^n \left[\frac{\partial\kappa_i}{\partial\rho} \left(\frac{\partial\rho}{\partial\alpha} \right)^2 + \kappa_i \frac{\partial^2\rho}{\partial\alpha^2} \right] = -\sum_{i=1}^n \left[k_i g^2 + \kappa_i \varrho \right],$$

with

$$\begin{aligned} k_i &:= k(\eta_i, \rho, y_i) := \frac{\partial\kappa_i}{\partial\rho} \\ &= y_i \frac{\partial^2\theta_i}{\partial\rho^2} - \left(\frac{\partial^2 b_i}{\partial\theta_i^2} \frac{\partial\theta_i}{\partial\rho} + \frac{\partial^2 b_i}{\partial\theta_i\partial\rho} \right) \frac{\partial\theta_i}{\partial\rho} - \frac{\partial b_i}{\partial\theta_i} \frac{\partial^2\theta_i}{\partial\rho^2} - \frac{\partial^2 b_i}{\partial\theta_i\partial\rho} \frac{\partial\theta_i}{\partial\rho} - \frac{\partial^2 b_i}{\partial\rho^2} + \frac{\partial^2 a_i}{\partial\rho^2} \\ &\overset{(3.2)}{=} \frac{y_i - \mu_i}{(\mu_i + \rho)^2} + \frac{\mu_i}{\rho(\mu_i + \rho)} + \mathrm{tri}(y_i + \rho) - \mathrm{tri}(\rho), \end{aligned}$$

where $\text{tri}(x) := \partial^2 \log \Gamma(x)/\partial x^2$ is the trigamma function, and

$$g := g(\alpha) := \frac{\partial \rho}{\partial \alpha}, \qquad \varrho := \varrho(\alpha) := \frac{\partial^2 \rho}{\partial \alpha^2}.$$

The ML estimator for the jth model solves the following constrained optimization problem

$$\begin{aligned} \max_{\beta,\alpha} \quad & \ell(\beta, \alpha) \\ \text{subject to} \quad & R_j^\top \beta_2 = 0. \end{aligned} \tag{3.3}$$

As a first step towards the solution, I construct the Lagrangian

$$L(\beta, \alpha, \nu_j) = \ell(\beta, \alpha) - \nu_j^\top (R_j^\top \beta_2),$$

where ν_j denotes the r_j-vector of Lagrange multipliers. Setting the first derivatives equal to zero yields the system of nonlinear equations

$$s_1(\beta, \alpha) = 0, \qquad s_2(\beta, \alpha) - R_j \nu_j = 0, \qquad s_\alpha(\beta, \alpha) = 0, \qquad R_j^\top \beta_2 = 0. \tag{3.4}$$

Following De Luca et al. (2018, p. 3 f.), I consider a one-step ML estimator that approximates the solution of the system. In contrast to iterative procedures such as Newton-Raphson, which are typically used for solving nonlinear equation systems, the one-step ML estimator admits closed-form expressions.

3.1 One-step ML estimator

In the remainder of the paper, I assume that all necessary conditions for the algebraic manipulations, e.g. rank conditions on the regressor matrix X, are satisfied. Detailed proofs are found in Appendix A.

I expand the estimating equations of (3.4) (except for $R_j^\top \beta_2 = 0$) around starting values $\bar{\beta} = (\bar{\beta}_1^\top, \bar{\beta}_2^\top)^\top$ and $\bar{\alpha}$. Further, $\bar{\rho} = \rho(\bar{\alpha})$, since the mapping from α to ρ is strictly monotonic (log-link). Using a first-order Taylor expansion and ignoring the remainder term yields

$$\begin{aligned} 0 &\approx \bar{s}_1 - \bar{H}_{11}(\beta_1 - \bar{\beta}_1) - \bar{H}_{12}(\beta_2 - \bar{\beta}_2) - \bar{H}_{1\alpha}(\alpha - \bar{\alpha}), \\ 0 &\approx \bar{s}_2 - \bar{H}_{21}(\beta_1 - \bar{\beta}_1) - \bar{H}_{22}(\beta_2 - \bar{\beta}_2) - \bar{H}_{2\alpha}(\alpha - \bar{\alpha}) - R_j \nu_j, \\ 0 &\approx \bar{s}_\alpha - \bar{H}_{\alpha 1}(\beta_1 - \bar{\beta}_1) - \bar{H}_{\alpha 2}(\beta_2 - \bar{\beta}_2) - \bar{H}_{\alpha\alpha}(\alpha - \bar{\alpha}), \\ 0 &= R_j^\top \beta_2, \end{aligned} \tag{3.5}$$

where $\bar{s}_p := s_p(\bar{\beta}, \bar{\alpha}), \bar{H}_{pq} := H_{pq}(\bar{\beta}, \bar{\alpha})$ and $\bar{H}_{p\alpha} = H_{p\alpha}(\bar{\beta}, \bar{\alpha})$, $p = 1, 2$. In the following, all quantities evaluated at $(\bar{\beta}, \bar{\alpha})$ are denoted by a bar and the approximations in (3.5) are treated as equalities for a simpler notation.

First, consider the unrestricted model with $R_u = 0$. Define the data transformations

$$\begin{aligned} \bar{y} &:= \bar{X}_1 \bar{\beta}_1 + \bar{X}_2 \bar{\beta}_2 + \bar{u}, & \bar{X}_p &:= \bar{\Psi}^{1/2} X_p, \quad p = 1, 2, \\ \bar{y}_0 &:= \bar{y} - \bar{g} \bar{\Psi}^{-1/2} \bar{C}(y - \bar{\mu}) \bar{\alpha}, & \bar{u} &:= \bar{\Psi}^{-1/2} \bar{V}(y - \bar{\mu}), \end{aligned} \tag{3.6}$$

which involve the $n \times n$ matrices

$$\begin{aligned} \bar{V} &:= V(\bar{\eta}, \bar{\rho}) := \operatorname{diag}\left(v(\bar{\eta}_1, \bar{\rho}), v(\bar{\eta}_2, \bar{\rho}), \ldots, v(\bar{\eta}_n, \bar{\rho})\right), \\ \bar{\Psi} &:= \Psi(\bar{\eta}, \bar{\rho}, y) := \operatorname{diag}\left(\psi(\bar{\eta}_1, \bar{\rho}, y_1), \psi(\bar{\eta}_2, \bar{\rho}, y_2), \ldots, \psi(\bar{\eta}_n, \bar{\rho}, y_n)\right), \\ \bar{C} &:= C(\bar{\eta}, \bar{\rho}) := \operatorname{diag}\left(c(\bar{\eta}_1, \bar{\rho}), c(\bar{\eta}_2, \bar{\rho}), \ldots, c(\bar{\eta}_n, \bar{\rho})\right), \end{aligned}$$

and the n-vectors $\bar{\mu} := \mu(\bar{\eta}) := (h(\bar{\eta}_1), h(\bar{\eta}_2), \ldots, h(\bar{\eta}_n))^\top$ and $\bar{\eta} := (\bar{\eta}_1, \bar{\eta}_2, \ldots, \bar{\eta}_n)^\top = X_1 \bar{\beta}_1 + X_2 \bar{\beta}_2$ with $\bar{\eta}_i := \eta(\bar{\beta}, x_i)$. Using the log-link further guarantees $\operatorname{rank}(\bar{\Psi}) = n$ because $\psi(\bar{\eta}_i, \bar{\rho}, y_i) = \bar{\mu}_i \bar{\rho}(y_i + \bar{\rho})/(\mu_i + \bar{\rho})^2 > 0$ since $\bar{\mu}_i > 0$, $\bar{\rho} > 0$ and $y_i \geq 0$ for all i. Moreover, define

$$\bar{t} := \bar{g} \bar{\kappa}^\top \mathbf{1} - \bar{g}(y - \bar{\mu})^\top \bar{C} \bar{\eta} - (\bar{g}^2 \bar{k}^\top \mathbf{1} + \bar{\varrho} \bar{\kappa}^\top \mathbf{1}) \bar{\alpha},$$

with n-vectors

$$\bar{k} := k(\bar{\eta}, \bar{\rho}, y) := (\bar{k}_1, \bar{k}_2, \ldots, \bar{k}_n)^\top, \quad \bar{\kappa} := \kappa(\bar{\eta}, \bar{\rho}, y) := (\bar{\kappa}_1, \bar{\kappa}_2, \ldots, \bar{\kappa}_n)^\top,$$

where $\bar{k}_i := k(\bar{\eta}_i, \bar{\rho}, y_i)$, $\bar{\kappa}_i := \kappa(\bar{\eta}_i, \bar{\rho}, y_i)$ and $\mathbf{1} := (1, \ldots, 1)^\top$ is an n-vector filled with ones. Notice the slight abuse in notation, where μ, k and κ are vector-valued functions here, whereas they were scalar-valued in the sections before. Furthermore, let

$$\bar{\epsilon} := \frac{\bar{g}}{\bar{g}^2 \bar{k}^\top \mathbf{1} + \bar{\varrho} \bar{\kappa}^\top \mathbf{1}}, \qquad \bar{q} := \bar{C}(y - \bar{\mu}).$$

Then, the solution to the linearized system of likelihood equations (3.5) can be expressed

in closed form as

$$
\begin{aligned}
\tilde{\beta}_{1u} = & \Bigg[\left(\frac{\bar{X}_1^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_1 \right)^{-1} \\
& + \Bigg\{ \left(\frac{\bar{X}_1^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_1 \right)^{-1} \left(\frac{\bar{X}_1^\top \bar{X}_2}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_2 \right) \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1} \\
& \cdot \left(\frac{\bar{X}_2^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_2^\top \bar{q}\bar{q}^\top X_1 \right) \left(\frac{\bar{X}_1^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_1 \right)^{-1} \Bigg\} \Bigg] \left(\frac{\bar{X}_1^\top \bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n} X_1^\top \bar{q} \right) \\
& - \Bigg[\left(\frac{\bar{X}_1^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_1 \right)^{-1} \left(\frac{\bar{X}_1^\top \bar{X}_2}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_2 \right) \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1} \\
& \cdot \left(\frac{\bar{X}_2^\top \bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n} X_2^\top \bar{q} \right) \Bigg], \\
\tilde{\beta}_{2u} = & - \Bigg[\left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1} \left(\frac{\bar{X}_2^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_2^\top \bar{q}\bar{q}^\top X_1 \right) \left(\frac{\bar{X}_1^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_1 \right)^{-1} \\
& \cdot \left(\frac{\bar{X}_1^\top \bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n} X_1^\top \bar{q} \right) \Bigg] + \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1} \left(\frac{\bar{X}_2^\top \bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n} X_2^\top \bar{q} \right), \\
\tilde{\alpha}_u = & - \frac{\bar{t} + \bar{g}(y - \bar{\mu})^\top \bar{C}(X_1 \tilde{\beta}_{1u} + X_2 \tilde{\beta}_{2u})}{\bar{g}^2 \bar{k}^\top \mathbf{1} + \bar{\varrho} \bar{\kappa}^\top \mathbf{1}},
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{M}_1 = & (I_n + \bar{g}\bar{\epsilon} \bar{\Psi}^{-1/2} \bar{q}\bar{q}^\top \bar{\Psi}^{-1/2}) \\
& - \left[(\bar{X}_1 + \bar{g}\bar{\epsilon} \bar{\Psi}^{-1/2} \bar{q}\bar{q}^\top X_1)(\bar{X}_1^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_1)^{-1} (\bar{X}_1^\top + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top \bar{\Psi}^{-1/2}) \right],
\end{aligned}
$$

is a symmetric matrix. In contrast to De Luca et al. (2018), $\bar{M}_1$ is not idempotent anymore due to the rank-1 perturbation in $I_n + \bar{g}\bar{\epsilon} \bar{\Psi}^{-1/2} \bar{q}\bar{q}^\top \bar{\Psi}^{-1/2}$, which is a consequence of the additional dispersion parameter ρ in the NB2 model compared to GLMs.

Likewise, consider the general one-step ML estimator for the jth model. Define the symmetric and idempotent $k_2 \times k_2$ matrix

$$
\bar{P}_j := \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2} R_j \left(R_j^\top \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1} R_j \right)^{-1} R_j^\top \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2},
$$

the $k_1 \times k_2$ matrix

$$
\bar{Q} := \left(\frac{\bar{X}_1^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_1 \right)^{-1} \left(\frac{\bar{X}_1^\top \bar{X}_2}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_2 \right) \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2},
$$

and the following transformation of the unrestricted one-step ML estimator $\tilde{\beta}_{2u}$

$$
\tilde{\vartheta} := \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{1/2} \tilde{\beta}_{2u}. \tag{3.7}
$$

Then, analogous to Proposition 1 of De Luca et al. (2018), I obtain the one-step ML estimator for the jth model in the following proposition.

Proposition 3.1 (One-step ML estimators)**.** *The one-step ML estimators of β_1, β_2 and α based on the jth model are*

$$\begin{aligned}\tilde{\beta}_{1j} &= \tilde{\beta}_{1r} - \bar{Q}\bar{W}_j\tilde{\vartheta},\\ \tilde{\beta}_{2j} &= \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n}\right)^{-1/2} \bar{W}_j\tilde{\vartheta},\\ \tilde{\alpha}_j &= -\frac{\bar{t} + \bar{g}(y-\bar{\mu})^\top \bar{C}(X_1\tilde{\beta}_{1j} + X_2\tilde{\beta}_{2j})}{\bar{g}^2\bar{k}^\top\mathbf{1} + \bar{\varrho}\bar{\kappa}^\top\mathbf{1}},\end{aligned}$$

where

$$\tilde{\beta}_{1r} = \left(\frac{\bar{X}_1^\top\bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n}X_1^\top\bar{q}\bar{q}^\top X_1\right)^{-1}\left(\frac{\bar{X}_1^\top\bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n}X_1^\top\bar{q}\right),$$

is the fully restricted one-step ML estimator of β_1 and $\bar{W}_j = I_{k_2} - \bar{P}_j$.

3.2 Transformed model

The WALS NB estimator relies on a preliminary transformation of the auxiliary regressors to reduce the computational burden, akin to WALS for the linear regression model (Magnus et al., 2010) and GLMs (De Luca et al., 2018).

First, scale the focus regressors by defining

$$\bar{Z}_1 := \bar{X}_1\bar{\Delta}_1, \qquad Z_1 := X_1\bar{\Delta}_1, \qquad \gamma_1 := \bar{\Delta}_1^{-1}\beta_1, \tag{3.8}$$

with the $k_1 \times k_1$ diagonal matrix $\bar{\Delta}_1 := \operatorname{diag}\left(\bar{X}_1^\top\bar{X}_1/n\right)^{-1/2}$ such that $\operatorname{diag}\left(\bar{Z}_1^\top\bar{Z}_1/n\right) = (1,\dots,1)$. The only purpose of the transformation is to improve the numerical accuracy by normalizing all regressors to be the same scale in $\bar{Z}_1$ (De Luca et al., 2018, p. 5). It further implies

$$\begin{aligned}\bar{Z}_1\gamma_1 &= \bar{X}_1\beta_1,\\ \beta_1 &= \bar{\Delta}_1\gamma_1,\\ \bar{M}_1 &= (I_n + \bar{g}\bar{\epsilon}\bar{\Psi}^{-1/2}\bar{q}\bar{q}^\top\bar{\Psi}^{-1/2})\\ &\quad - \left[(\bar{Z}_1 + \bar{g}\bar{\epsilon}\bar{\Psi}^{-1/2}\bar{q}\bar{q}^\top Z_1)(\bar{Z}_1^\top\bar{Z}_1 + \bar{g}\bar{\epsilon}Z_1^\top\bar{q}\bar{q}^\top Z_1)^{-1}(\bar{Z}_1^\top + \bar{g}\bar{\epsilon}Z_1^\top\bar{q}\bar{q}^\top\bar{\Psi}^{-1/2})\right],\\ &\overset{(3.8)}{=} (I_n + \bar{g}\bar{\epsilon}\bar{\Psi}^{-1/2}\bar{q}\bar{q}^\top\bar{\Psi}^{-1/2})\\ &\quad - \left[(\bar{X}_1 + \bar{g}\bar{\epsilon}\bar{\Psi}^{-1/2}\bar{q}\bar{q}^\top X_1)(\bar{X}_1^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1)^{-1}(\bar{X}_1^\top + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top\bar{\Psi}^{-1/2})\right],\end{aligned}$$

so scaling by $\bar{\Delta}_1$ has no effect on $\bar{M}_1$. Next, transform the auxiliary regressors by

$$\bar{Z}_2 := \bar{X}_2\bar{\Delta}_2\bar{\Xi}^{-1/2}, \qquad Z_2 := X_2\bar{\Delta}_2\bar{\Xi}^{-1/2}, \qquad \gamma_2 := \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\beta_2, \tag{3.9}$$

where

$$\bar{\Xi} := \frac{\bar{\Delta}_2\bar{X}_2^\top\bar{M}_1\bar{X}_2\bar{\Delta}_2}{n}, \tag{3.10}$$

and I assumed $\bar{X}_2^\top \bar{M}_1 \bar{X}_2$ to be positive definite so $\bar{\Xi}^{1/2}$ exists. Furthermore, the $k_2 \times k_2$ diagonal matrix $\bar{\Delta}_2 := \operatorname{diag}\left(\bar{X}_2^\top \bar{M}_1 \bar{X}_2/n\right)^{-1/2}$ is chosen such that $\operatorname{diag}(\bar{\Xi}) = (1, \dots, 1)$. Unlike the matrix $\bar{\Delta}_1$, the transformation by $\bar{\Delta}_2$ serves the dual purpose of improving numerical accuracy and making the WALS NB estimator equivariant to scale transformations of the auxiliary regressors. Otherwise it would be only scale equivariant for the focus regressors (De Luca and Magnus, 2011, p. 528).

Notice that combining (3.9) and (3.10) leads to

$$\frac{\bar{Z}_2^\top \bar{M}_1 \bar{Z}_2}{n} = I_{k_2}. \tag{3.11}$$

In contrast to De Luca et al. (2018), $\bar{M}_1 \bar{Z}_2/\sqrt{n}$ is not semiorthogonal[1] anymore, since $\bar{M}_1$ is not idempotent. The transformation further implies

$$\begin{aligned} \bar{Z}_2 \gamma_2 &= \bar{X}_2 \beta_2, \\ \beta_2 &= \bar{\Delta}_2 \bar{\Xi}^{-1/2} \gamma_2. \end{aligned}$$

Using (3.8) and (3.9) I can show for the unrestricted model that

$$\eta = Z_1 \gamma_1 + Z_2 \gamma_2 = X_1 \beta_1 + X_2 \beta_2,$$

so the linear predictor stays the same for the unrestricted model. Therefore, all the quantities that only depend on $\bar{\alpha}$ and indirectly on $\bar{\beta}$ via

$$\bar{\eta} = X_1 \bar{\beta}_1 + X_2 \bar{\beta}_2 = Z_1 \bar{\gamma}_1 + Z_2 \bar{\gamma}_2,$$

where $\bar{\gamma}_1 = \bar{\Delta}_1^{-1} \bar{\beta}_1$ and $\bar{\gamma}_2 = \bar{\Xi}^{1/2} \bar{\Delta}_2^{-1} \bar{\beta}_2$, remain the same (e.g. $\bar{\mu}$, $\bar{\Psi}$, $\bar{Q}$, $\bar{g}$, $\bar{t}$, $\dots$), as they do not depend on $\bar{\beta}$ directly. Note that De Luca et al. (2018, p. 6) suggest using the fully iterated unrestricted ML estimates as starting values $\bar{\beta}$ and $\bar{\alpha}$. In this case, the starting value of the dispersion parameter $\bar{\alpha}_Z$ for the transformed regressors Z is identical to $\bar{\alpha}$ for the original regressors X since the estimated conditional means are equal, i.e. $h(\eta(\bar{\beta}, x_i)) = h(\eta(\bar{\gamma}, z_i))$ for all i, where $z_i^\top$ is the ith row vector of $Z = (Z_1, Z_2)$.

3.3 One-step ML estimation of transformed models

It follows from Proposition 3.1 using (3.11) that the one-step ML estimators for the jth transformed model are given by

$$\begin{aligned} \tilde{\gamma}_{1j} &= \tilde{\gamma}_{1r} - \bar{D} W_j \tilde{\gamma}_{2u}, \\ \tilde{\gamma}_{2j} &= \left(\frac{\bar{Z}_2^\top \bar{M}_1 \bar{Z}_2}{n} \right)^{-1/2} \bar{W}_{Z,j} \tilde{\vartheta}_Z = W_j \tilde{\gamma}_{2u}, \\ \tilde{\alpha}_j &= -\frac{\bar{t} + \bar{g}(y - \bar{\mu})^\top \bar{C}(Z_1 \tilde{\gamma}_{1j} + Z_2 \tilde{\gamma}_{2j})}{\bar{g}^2 \bar{k}^\top \mathbf{1} + \bar{\varrho} \bar{\kappa}^\top \mathbf{1}}, \end{aligned} \tag{3.12}$$

[1] Semiorthogonality is defined as $AA^\top = I$ or $A^\top A = I$ for a general (non-square) matrix A (Zhang, 2017, p. 104).

where the fully restricted and unrestricted estimators are

$$\begin{aligned}
\tilde{\gamma}_{1r} &= \left(\frac{\bar{Z}_1^\top \bar{Z}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} Z_1^\top \bar{q}\bar{q}^\top Z_1\right)^{-1} \left(\frac{\bar{Z}_1^\top \bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n} Z_1^\top \bar{q}\right) = \bar{\Delta}_1^{-1} \tilde{\beta}_{1r}, \\
\tilde{\gamma}_{1u} &= \tilde{\gamma}_{1r} - \bar{Q}_Z \tilde{\gamma}_{2u} = \bar{\Delta}_1^{-1} \left[\tilde{\beta}_{1r} - \bar{Q}\tilde{\vartheta}\right] = \bar{\Delta}_1^{-1} \tilde{\beta}_{1u}, \\
\tilde{\gamma}_{2u} &= -\left(\frac{\bar{Z}_2^\top \bar{Z}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} Z_2^\top \bar{q}\bar{q}^\top Z_1\right) \left(\frac{\bar{Z}_1^\top \bar{Z}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} Z_1^\top \bar{q}\bar{q}^\top Z_1\right)^{-1} \left(\frac{\bar{Z}_1^\top \bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n} Z_1^\top \bar{q}\right) \\
&\quad + \left(\frac{\bar{Z}_2^\top \bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n} Z_2^\top \bar{q}\right) \\
&= \bar{\Xi}^{1/2} \bar{\Delta}_2^{-1} \tilde{\beta}_{2u},
\end{aligned} \tag{3.13}$$

with

$$\bar{Q}_Z = \left(\frac{\bar{Z}_1^\top \bar{Z}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} Z_1^\top \bar{q}\bar{q}^\top Z_1\right)^{-1} \left(\frac{\bar{Z}_1^\top \bar{Z}_2}{n} + \frac{\bar{g}\bar{\epsilon}}{n} Z_1^\top \bar{q}\bar{q}^\top Z_2\right) := \bar{D}.$$

Exploiting $R_j^\top R_j = I_{r_j}$, the following terms simplify

$$\begin{aligned}
\bar{P}_{Z,j} &= R_j (R_j^\top R_j)^{-1} R_j^\top = R_j R_j^\top =: P_j, \\
\bar{W}_{Z,j} &= I_{k_2} - \bar{P}_{Z,j} = I_{k_2} - P_j =: W_j.
\end{aligned}$$

Analogous to $\tilde{\vartheta}$, using (3.11) yields

$$\tilde{\vartheta}_Z = \left(\frac{\bar{Z}_2^\top \bar{M}_1 \bar{Z}_2}{n}\right)^{1/2} \tilde{\gamma}_{2u} = \tilde{\gamma}_{2u}.$$

As a direct consequence of (3.11), both P_j and W_j become nonrandom projection matrices that are different from $\bar{P}_j$ and $\bar{W}_j$ used for the estimation with the untransformed regressors. Furthermore, W_j reduces to a diagonal matrix with $k_2 - r_j$ ones and r_j zeros on its main diagonal. The hth diagonal element of W_j is zero, when the hth component of γ_2 is constrained to be zero in the jth model. Otherwise, the hth component is one. Combining this observation with $\tilde{\gamma}_{2j}$ from (3.12), it follows that all models that include the hth column of Z_2 as regressor will have the same estimator for the hth component, namely the hth component of $\tilde{\gamma}_{2u}$.

Note that the jth model for the transformed regressors is generally not equivalent to the jth model of the untransformed regressors because the restriction in (3.4) differs. The exceptions are the unrestricted model u and the fully restricted model r, where the restriction is irrelevant:

1. For the unrestricted model, the restriction matrix is zero, i.e. $R_u = 0$.
2. For the fully restricted model, the restriction matrix is the identity matrix, i.e. $R_r = I_{k_2}$. However, this is equivalent to estimating an unrestricted model containing only the focus regressors.

This implies that $\tilde{\gamma}_{2j} \neq \bar{\Xi}^{1/2} \bar{\Delta}_2^{-1} \tilde{\beta}_{2j}$ for $j \notin \{u, r\}$ and $k_2 \geq 2$ auxiliary regressors. For $k_2 = 1$, there exist only two models: 1. the unrestricted and 2. the fully restricted model, so $j \in \{u, r\}$. The results are summarized in Corollary 3.2.

Corollary 3.2. *Let u and r be the indices that denote the unrestricted and fully restricted estimators with $R_u = 0$ and $R_r = I_{k_2}$, respectively. Then, for $j \neq \{u, r\}$ and $k_2 \geq 2$:*

$$\tilde{\gamma}_{1j} \neq \bar{\Delta}_1^{-1}\tilde{\beta}_{1j}, \qquad \tilde{\gamma}_{2j} \neq \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\tilde{\beta}_{2j}.$$

For $k_2 = 1$, either $j = u$ or $j = r$ holds, so the general relationships for the unrestricted and fully restricted estimators apply:

$$\tilde{\gamma}_{1u} = \bar{\Delta}_1^{-1}\tilde{\beta}_{1u}, \qquad \tilde{\gamma}_{2u} = \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\tilde{\beta}_{2u},$$

and

$$\tilde{\gamma}_{1r} = \bar{\Delta}_1^{-1}\tilde{\beta}_{1r}, \qquad \tilde{\gamma}_{2r} = 0.$$

4 WALS NB model averaging estimator

Consider the model averaging estimators of γ_1, γ_2 and α

$$\hat{\gamma}_1 = \sum_{j=1}^{2^{k_2}} \lambda_j \tilde{\gamma}_{1j}, \qquad \hat{\gamma}_2 = \sum_{j=1}^{2^{k_2}} \lambda_j \tilde{\gamma}_{2j}, \qquad \hat{\alpha} = \sum_{j=1}^{2^{k_2}} \lambda_j \tilde{\alpha}_j,$$

where λ_j are data-dependent model weights satisfying the restrictions

$$0 \leq \lambda_j \leq 1, \qquad \sum_{j=1}^{2^{k_2}} \lambda_j = 1, \qquad \lambda_j = \lambda_j(\sqrt{n}\tilde{\gamma}_{2u}). \tag{4.1}$$

Note that the regularity condition $\lambda_j = \lambda_j(\sqrt{n}\tilde{\gamma}_{2u})$ is equivalent to the condition on the model weights used by Hjort and Claeskens (2003).

From (3.12) I get

$$\begin{aligned} \hat{\gamma}_1 &= \tilde{\gamma}_{1r} - \bar{D}W\tilde{\gamma}_{2u} = \tilde{\gamma}_{1r} - \bar{D}\hat{\gamma}_2, \\ \hat{\gamma}_2 &= W\tilde{\gamma}_{2u}, \\ \hat{\alpha} &= -\frac{\bar{t} + \bar{g}(y - \bar{\mu})^\top \bar{C}(Z_1\hat{\gamma}_1 + Z_2\hat{\gamma}_2)}{\bar{g}^2\bar{k}^\top\mathbf{1} + \bar{\varrho}\bar{\kappa}^\top\mathbf{1}}, \end{aligned} \tag{4.2}$$

where $W = \sum_{j=1}^{2^{k_2}} \lambda_j W_j$ is a diagonal matrix with entries $w_h \in [0, 1]$, because W_j is a diagonal matrix with entries $w_{j,h} \in \{0, 1\}$, $h = 1, 2, \ldots, k_2$ (notice the slight abuse of notation: h is used as an index here and does not refer to the inverse link). Next, I can transform $\hat{\alpha}$ to an estimate for ρ by applying the inverse of the log-link, i.e. $\hat{\rho} = \exp(\hat{\alpha})$. Furthermore, using $\hat{\gamma}_1$ and $\hat{\gamma}_2$, the WALS estimators of the original parameters β_1 and β_2 are given by

$$\hat{\beta}_1 = \bar{\Delta}_1\hat{\gamma}_1, \qquad \hat{\beta}_2 = \bar{\Delta}_2\bar{\Xi}^{-1/2}\hat{\gamma}_2.$$

The final step in completing the WALS NB model averaging estimator is to estimate the model weights λ_j. However, notice that both $\hat{\gamma}_1$ and $\hat{\alpha}$ can be expressed as functions of $\hat{\gamma}_2$. Therefore, it is sufficient to find an expression for $\hat{\gamma}_2$ instead of directly estimating the

weights λ_j. Similar to De Luca et al. (2018, p. 6), I construct $\hat{\gamma}_2$ as a Bayesian shrinkage estimator by exploiting the approximate normality and independence of $\tilde{\gamma}_{2u}$ under the local misspecification framework (see e.g. Hjort and Claeskens, 2003). First, let the auxiliary parameters be $\beta_2 = \delta/\sqrt{n}$, where δ is an unknown constant vector that represents the departure of the DGP from the unrestricted model. Then, if the fully iterated ML estimator of the unrestricted model is used as starting values $\bar{\beta}_1$, $\bar{\beta}_2$ and $\bar{\alpha}$ and mild regularity conditions are assumed, I can show that

$$\sqrt{n}\tilde{\gamma}_{2u} \approx \mathcal{N}(\sqrt{n}\gamma_{2n}, I_{k_2}) = \mathcal{N}(d, I_{k_2}), \tag{4.3}$$

in large samples, where $\gamma_{2n} = d/\sqrt{n}$, $d = \Xi^{1/2}\Delta_2^{-1}\delta$, $\Xi = \operatorname{plim}\bar{\Xi}$ and $\Delta_2 = \operatorname{plim}\bar{\Delta}_2$ (see the supplementary materials for more details). Further, consider $\hat{\gamma}_2$ from (4.2) and assume analogously to De Luca et al. (2018, p. 6) that each diagonal element $w_h, h = 1, 2, \ldots, k_2$, of W only depends on the hth component $\sqrt{n}\tilde{\gamma}_{2u,h}$ of $\sqrt{n}\tilde{\gamma}_{2u}$. Then, (4.3) implies that the components of $\hat{\gamma}_2$ are also approximately independent. This assumption further simplifies the estimation problem by reducing the k_2-dimensional problem of estimating $\hat{\gamma}_2$ to k_2 times a one-dimensional problem of estimating each element of $\hat{\gamma}_2$. Moreover, $\hat{\gamma}_{2,h}$ is a shrunken version of $\tilde{\gamma}_{2u,h}$ because $0 \leq w_h \leq 1$, therefore, $\hat{\gamma}_2$ is a shrinkage estimator of γ_{2n}.

The previous two observations suggest that the Bayesian posterior mean is a suitable shrinkage estimator for $\sqrt{n}\gamma_{2n,h}$. Thus, the hth component of the WALS NB estimator $\hat{\gamma}_2$ follows as

$$\hat{\gamma}_{2,h} = \frac{\mathsf{E}(\sqrt{n}\gamma_{2n,h}|\sqrt{n}\tilde{\gamma}_{2u,h})}{\sqrt{n}} = \frac{\mathsf{E}(d_h|\sqrt{n}\tilde{\gamma}_{2u,h})}{\sqrt{n}}, \quad d_h \sim f, \tag{4.4}$$

where $\sqrt{n}\tilde{\gamma}_{2u,h} \approx \mathcal{N}(d_h, 1)$ with prior mean d_h, which is the hth element of d and is assumed to have a symmetric and unimodal prior f (see section 9 of Magnus and De Luca (2016) for more details on the prior and the estimation). Notice again that $\hat{\gamma}_{2,h}$ lies between 0 and the 'observed data' $\tilde{\gamma}_{2u,h}$.

Magnus and De Luca (2016) require the desirable properties of robustness[2], neutrality[3] and minimax regret[4] for the prior f, which further motivates the use of the Bayesian posterior mean as the shrinkage estimator in $\hat{\gamma}_{2,h}$. The reflected Weibull, under suitable parameter values, is a prior that fulfills all the properties mentioned above. In contrast, the Laplace prior is neutral but not robust (Magnus and De Luca, 2016, p. 132). However, it admits a closed-form expression for the posterior mean in (4.4) (see e.g. Theorem 1 in Magnus et al., 2010) and therefore calculating the posterior mean under the Laplace prior is computationally less complex than under the reflected Weibull, which requires numerical integration.

[2]A prior $\pi(\gamma)$ is robust if the posterior mean $m(x)$ based on π satisfies $x - m(x) \to 0$ as $x \to \infty$.

[3]A prior $\pi(\gamma)$ is neutral if the prior median of γ is zero and the prior median of $|\gamma|$ is one.

[4]Regret is defined as difference between risk and the infimum of risk, where risk is defined as expected squared loss.

5 Performance metrics

In order to compare the performance of WALS NB with other methods, I first need to define performance metrics. The classical performance measure for regression is the RMSE, which is given by

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_i - y_i)^2}, \tag{5.1}$$

where $\hat{\mu}_i$ is the predicted mean for observation i. However, I would like to evaluate the fit of an entire distribution and not only the expectation. Traditional measures used in machine learning such as (R)MSE only focus on point predictions, i.e. the conditional expectation of the fitted distribution, in relation to the observed values and do not make judgment on other aspects of the fitted distribution. Czado et al. (2009) recommend scoring rules for evaluation of count data models, which have also been used in Kolassa (2016). WALS NB and all other methods considered in this paper fit an entire (conditional) distribution for each individual that allows probabilistic predictions/forecasts, which is exactly the scenario for which scoring rules provide quality assessment (Gneiting and Raftery, 2007, p. 359). For count data, a probabilistic forecast is a predictive probability distribution $\hat{P}$ on the set of nonnegative integers $\mathbb{N}_0$ (Czado et al., 2009, p. 1254).

Following Czado et al. (2009, p. 1256), I take scoring rules to be penalties I wish to minimize. Specifically, the penalty $s(\hat{P}, y)$ is incurred when the forecaster quotes predictive distribution $\hat{P}$ and count y is realized. Moreover, let $s(\hat{P}, Q)$ denote the expected value of $s(\hat{P}, \cdot)$ under distribution Q

$$s(\hat{P}, Q) = \int s(\hat{P}, y) dQ(y).$$

In practice, the average over suitable pairs $(\hat{P}, y)$ is used:

$$S := \frac{1}{n}\sum_{i=1}^{n} s(\hat{P}_i, y_i), \tag{5.2}$$

where $\hat{P}_i$ refers to the ith predictive distribution and y_i the ith observed count. In the simulation experiment and empirical application of Sections 6 and 7, respectively, scores will always refer to a suitable average.

Suppose the forecaster has predictive distribution Q available. Then the forecaster has no incentive to predict any $\hat{P} \neq Q$ and is encouraged to quote her true belief, $\hat{P} = Q$, if the scoring rule is *strictly proper*. Strict propriety is defined by

$$s(Q, Q) \leq s(\hat{P}, Q),$$

with equality if and only if $\hat{P} = Q$, and encourages honest quotes (Czado et al., 2009, p. 1256; Gneiting and Raftery, 2007, p. 360). If $s(Q, Q) \leq s(\hat{P}, Q)$ for all $\hat{P}$ and Q, then the scoring rule is only proper. Since only strict propriety ensures that both calibration (consistency with actual realizations) and sharpness (concentration of the predictive distribution) of

the predictive distribution are addressed (Winkler, 1996), I exclusively use strictly proper scoring rules.

Czado et al. (2009, p. 1256 f.) propose a number of strictly proper scoring rules for count data. It is a priori unclear which scoring rule to use unless there is a unique and clearly defined underlying decision problem. Since probabilistic forecasts often have many uses, it is appropriate to use a variety of scores to take advantage of their differing emphases (Czado et al., 2009, p. 1257). In this paper, I use the logarithmic (log), Brier and spherical score, which I briefly summarize here: Let $\hat{p}_y := \hat{P}(Y = y)$ denote the probability mass at count y (for continuous distributions it is the density at y), then the *log score* is defined as

$$\text{logs}(\hat{P}, y) = -\log(\hat{p}_y). \tag{5.3}$$

The sum of log scores corresponds to the negative log-likelihood. Further define

$$||\hat{p}||^2 = \sum_{r=0}^{\infty} \hat{p}_r^2, \tag{5.4}$$

where the infinite sum may be truncated if no closed-form expression exists. The *quadratic score*, also called *Brier score*, is then

$$\text{qs}(\hat{P}, y) = -2\hat{p}_y + ||\hat{p}||^2. \tag{5.5}$$

The *spherical score* uses the same components differently:

$$\text{sphs}(\hat{P}, y) = -\frac{\hat{p}_y}{||\hat{p}||}. \tag{5.6}$$

6 Simulation experiment

The aim is to compare the performance of WALS NB with the traditional ML estimator of the NB2 regression model in a controlled environment. The DGP is inspired by the local misspecification framework so I can assess the influence of varying numbers of focus and auxiliary regressors.

The dependent count variable is sampled from an NB2 using a log-link, i.e.

$$\begin{aligned} y_i|x_i &\sim \text{NB2}(\mu_i, \rho), \\ \mu_i &= \exp(\alpha + x_i^\top \beta), \\ x_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_k), \end{aligned} \tag{6.1}$$

for $i = 1, 2, \ldots, n$, where $x_i = (x_{i1}^\top, x_{i2}^\top)^\top$ is a random vector of dimension $k = k_1 + k_2$ composed of k_1 focus regressors x_{i1} and k_2 auxiliary regressors x_{i2}. Analogously, the coefficient vector is separated into two parts: $\beta = (\beta_1^\top, \beta_2^\top)^\top$. Inspired by the simulation experiments in Zhang and Liu (2019) and De Luca et al. (2023), who compare confidence and prediction intervals of model averaging methods for the linear regression model, I choose the regressors to be multivariate normal because it allows me to analyze the effect of

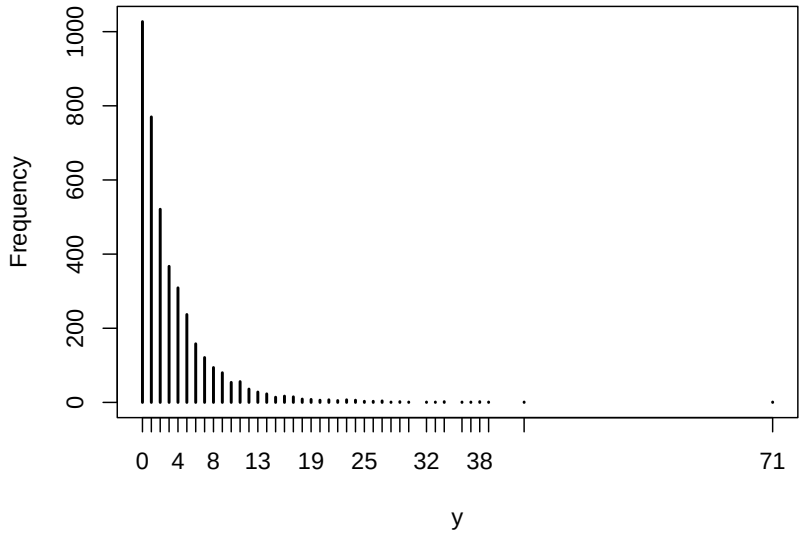

Figure 1: Visualization of count variable y in the training set of the first simulation run of setting $k_1 = 10, k_2 = 100, \rho = 1, b = 0$ with $n = 4000$ observations.

the correlation between the regressors on the performance of the methods. For simplicity, I specify each element of x_i to have variance 1 and pairwise correlation b, i.e.

$$\Sigma_k = \begin{pmatrix} 1 & b & \cdots & b \\ b & 1 & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \cdots & 1 \end{pmatrix}.$$

The same offset $\alpha = \log(3)$ is used in all experiments such that the DGP produces reasonable counts, see Figure 1 for a visualization of a training set from a specific run.

Moreover, the regression coefficients are generated as follows: Define the vectors $\bar{\beta}_1 := (\bar{\beta}_{1,1}, \bar{\beta}_{1,2}, \ldots, \bar{\beta}_{1,10})^\top$ and $\bar{\beta}_2 := (\bar{\beta}_{2,1}, \bar{\beta}_{2,2}, \ldots, \bar{\beta}_{2,100})^\top$. Then, the maximum number of regression coefficients $k_1 = 10$ and $k_2 = 100$ are randomly sampled *once* according to the following rules:

- $\bar{\beta}_1$: Each element $\bar{\beta}_{1,j}, j = 1, 2, \ldots, 10$, is drawn independently with 50% chance from $U(-0.25, -0.1)$ or from $U(0.1, 0.25)$, so that both positive and negative coefficients are present.
- $\bar{\beta}_2$: Each element is drawn independently from a uniform, i.e.

$$\bar{\beta}_{2,m} \sim U(-0.01, 0.01), \quad m = 1, 2, \ldots, 100.$$

The simulations then only take the first k_1 and k_2 values from these vectors as regression coefficients β_1 and β_2. For example, in the setting $k_1 = 5, k_2 = 10$, $\beta_1 = (\bar{\beta}_{1,1}, \bar{\beta}_{1,2}, \ldots, \bar{\beta}_{1,5})^\top$ and $\beta_2 = (\bar{\beta}_{2,1}, \bar{\beta}_{2,2}, \ldots, \bar{\beta}_{2,10})^\top$. Hence, the magnitude of the elements in β_2 is much smaller than in β_1, therefore the regressors x_{i2} are considered auxiliary regressors and the main variation is driven by x_{i1}. Table C.1 and C.2 in Appendix C show the entries of $\bar{\beta}_1$

Table 1: Simulation design – Parameters and choice of values

Parameter	Description	Values
k_1	# focus regressors	1, 5, 10
k_2	# auxiliary regressors	1, 5, 10, 20, 50, 100
n	# training observations	500, 1000, 2000, 3000, 4000
b	correlation coefficient of regressors	0, 0.3, 0.5, 0.9
α	constant	log(3)
ρ	dispersion parameter	1, 1.5, 2
R	# runs	300

– Each simulation scenario is a unique combination of the parameter values, which produces 1080 scenarios in total.

and $\bar{\beta}_2$, respectively.

All values of the parameters used in the experiment are summarized in Table 1. A total of 1080 scenarios consisting of all combinations of the parameters are simulated for $R = 300$ runs each.

I compare six different procedures that are named according to the pattern 'method-specification'. The two methods are called 'walsNB', which estimates the NB2 regression model using WALS NB, and 'ML', which uses maximum likelihood. For WALS NB procedures, the Weibull prior is used as it theoretically provides the best tradeoff between robustness and regret, for more details see Magnus and De Luca (2016, p. 130 ff.). The results for other priors are expected to be quite similar as WALS for the linear regression model has empirically shown to be relatively insensitive to the choice of the prior (De Luca et al., 2022).[5]

The procedures considered are

1. walsNB-dgp: Emulates the DGP (6.1) by including x_{i1} and a constant as focus regressors and x_{i2} as auxiliary regressors.

2. walsNB-aux: Includes only a constant as focus regressor and the 'true' regressors x_i as auxiliary.

3. ML-U: Includes a constant and the 'true' regressors x_i.

4. ML-focus: Only includes the focus regressors x_{i1} and a constant.

5. ML-AC: Only includes the auxiliary regressors x_{i2} and a constant.

6. oracle: The true model of the DGP (6.1) that is not estimated.

The second WALS NB specification, walsNB-aux, is included to analyze the extent to which prior information about the focus regressors in walsNB-dgp affects performance. Ideally, including x_{i1} as focus regressors in the procedure should improve performance as they are the covariates that dominate and should therefore be included in all submodels of

[5] I also conducted the simulation experiment using the Laplace prior and the results are similar to the ones using the Weibull prior.

WALS NB. However, in walsNB-aux their coefficients are also subjected to the regularization of the Bayesian estimation step, which may improve performance. Thus, a priori it is unclear which model will dominate.

All ML specifications are estimated using a log-link for the mean parameter, while the dispersion parameter is estimated directly without a link (default setting). Moreover, I increase the maximum number of iterations for both the alternation process between IRLS and ML estimation of ρ and the IRLS algorithm itself from the default setting of 25 to 2500 to increase the odds for convergence. The remaining settings, e.g. convergence criteria, are left at their default values.

Moreover, the WALS NB specifications use a log-link for the mean and dispersion parameter following the DGP (6.1) and are initialized using the ML estimates of the unrestricted model, which are given by the ML-U procedure (using the increased maximum number of iterations as described above). This initialization is recommended by De Luca et al. (2018, p. 4) as it produces lower RMSE for the WALS GLM estimator in their Monte Carlo simulations (see Table 3 in De Luca et al., 2018, p. 11) compared to using the estimates of the fully restricted model as starting values. It further ensures that (4.3) approximately holds for the one-step estimators of the auxiliary regression coefficients, which I exploit in the Bayesian estimation step to reduce the k_2-dimensional posterior mean estimation to k_2 one-dimensional problems (see Section 4).

Finally, the Weibull prior for all WALS specifications uses the parameters recommended in Magnus and De Luca (2016, p. 132), which are minimax regret solutions for the normal location problem.

In order to compare the performance of the procedures, I follow the benchmark experiments framework of Hothorn et al. (2005, p. 681 f.) and more specifically the 'Simulation Problem'. The simulation is structured to emulate the typical use of the methods: For each scenario and run, a training sample of size n is generated, where all procedures are applied and performance criteria are computed on an independently generated validation set that is fixed in size to $n_e = 4000$ to avoid any variation due to its size. I do not employ hypothesis tests to check if the performance differences are significant because the simulation experiment itself is already computationally intensive due to the large number of parameter settings.

I consider the RMSE as a classical precision measure for regression and additionally log, Brier and spherical scores to assess the distributional fit as described in Section 5. For the scoring rules, the average is taken over the validation sample as in (5.2). Further, I truncate the infinite sum in $||\hat{p}||$ from (5.4) used in the Brier and spherical score at the count $r = 150$ because the response typically does not exceed 150 and it would be meaningless to extrapolate beyond the observed data. See Figure 1 for a visualization of the training data of a single simulation run of the setting $k_1 = 10, k_2 = 100, \rho = 1, b = 0$ (this setting should maximize the range of y, since $k_1 = 10$ and $k_2 = 100$ allow for the largest possible means μ_i and $\rho = 1$ maximizes variance), where the response only ranges between 0 and 71.

In the following, only the scenarios highlighting the differences between the procedures are discussed. For more results, see the supplementary materials.

6.1 Varying the number of regressors

First, I analyze how the procedures behave when the number of regressors is varied. In the following plots, the points represent the mean validation metric over all successful runs of the experiment, i.e. $R = 300$ if the method never fails to converge. The total number of failed runs is given below the corresponding points in the plots and the shaded area displays the interquartile range (the box of a boxplot) of the validation metric.

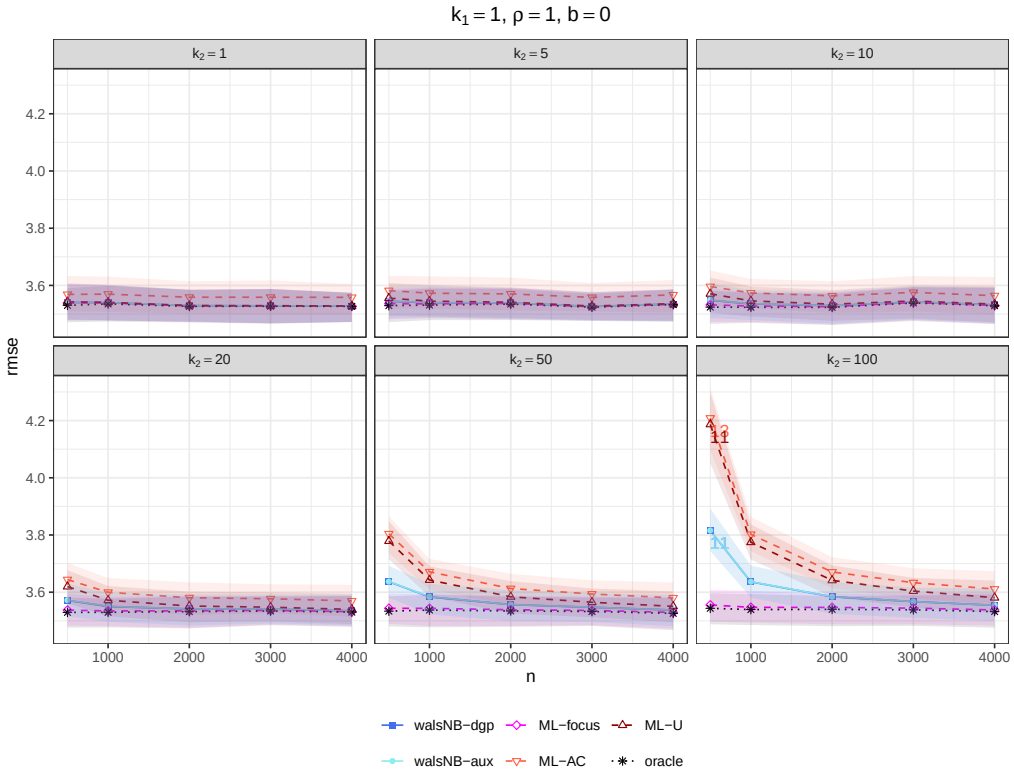

Figure 2: Mean validation RMSE and quartiles varying n and k_2
The remaining parameters are fixed at $k_1 = 1, \rho = 1$ and $b = 0$. The shaded areas show the interquartile range. The number below a point indicates how often the method failed to converge in this particular setting.

Figure 2 shows that walsNB-aux performs similarly to walsNB-dgp in terms of mean validation RMSE when we vary k_2 with fixed $k_1 = 1$, $\rho = 1$ and $b = 0$, because most of the regressors are auxiliary and the former includes all regressors as auxiliary. Moreover, both WALS NB specifications outperform ML-U on average when $k_2 \geq 20$ and n is small ($n \leq 2000$). In fact, WALS NB specifications show lower mean validation RMSE than all ML specifications in these scenarios, except for ML-focus that contains only the focus regressors. For $k_2 = 100$ and $n < 2000$ the 'typical' performance of walsNB-dgp and walsNB-aux is also better than ML-U and ML-AC as their interquartile ranges do not overlap. On the other hand, when k_2 is small and/or n is large, their interquartile ranges are similar. The largest difference in mean RMSE is observed at $n = 500$, $k_1 = 1$, $k_2 = 100$ where walsNB-aux exhibits around 8.9% lower mean RMSE than ML-U. ML-focus performs the best in all scenarios, especially when k_2 is large and n small.

Therefore, if we know the focus regressors, then ML-focus yields the best fit in very

sparse situations with few observations. Otherwise, walsNB-dgp and walsNB-aux are better than using all regressors in the large regression model ML-U. The outperformance in walsNB-dgp and walsNB-aux compared to ML-U is likely due to the reduced variance thanks to the Bayesian regularization step, which typically reduces variance and leads to lower RMSE via the bias-variance trade-off. In reality it is unlikely that we can exactly identify which regressors are the focus regressors, so walsNB-aux offers a great alternative that does not require variable selection. In all scenarios, it performs at least as well as ML-U but better when the data is sparse and few observations are available. For large n or small k_2, all procedures fit the data equally well as their mean validation RMSE converges to the RMSE of the oracle.

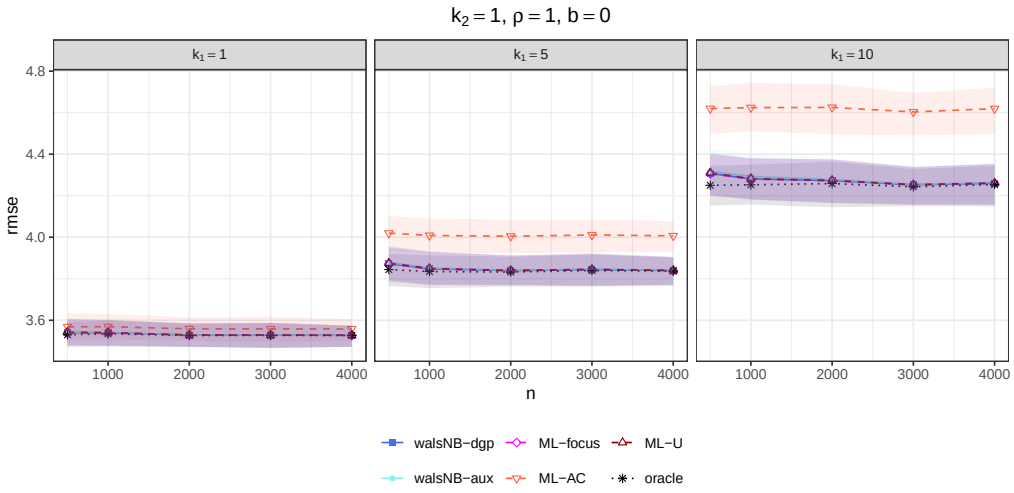

Figure 3: Mean validation RMSE and quartiles varying n and k_1
The remaining parameters are fixed at $k_2 = 1, \rho = 1$ and $b = 0$. The shaded areas show the interquartile range.

When I vary k_1 with fixed $k_2 = 1$, $\rho = 1$ and $b = 0$ in Figure 3, the picture changes. In all scenarios, ML-AC returns the highest RMSE and the remaining specifications perform similarly as their interquartile ranges overlap. Increasing k_1 shifts the 'RMSE-curve' up for all procedures, including the oracle, while retaining their relative order. This behavior is explained by the form of the variance of the NB2 distribution in (2.2). The more focus regressors with large coefficients are included, the more likely it is that the conditional expectation μ_i is large, which increases the variation of the response y_i since the conditional variance is monotonically increasing in μ_i. Thus, even if we could exactly estimate the true β_1 and β_2, the RMSE would still increase due to the increased conditional variance, which is demonstrated by the behavior of the oracle.

The same patterns hold for the validation log score. Firstly, Figure 4 shows that WALS NB specifications generally perform better than ML specifications in terms of log score, when the number of auxiliary regressors is high compared to the number of focus regressors and few observations are available. The exception is again ML-focus, which performs the best across all scenarios. The largest difference in mean log score between walsNB-aux and ML-U is realized at $k_1 = 1$, $k_2 = 100$ and $n = 500$ where the mean log score of walsNB-aux is around 3.9% lower. Moreover, the typical performance of walsNB-aux is also better

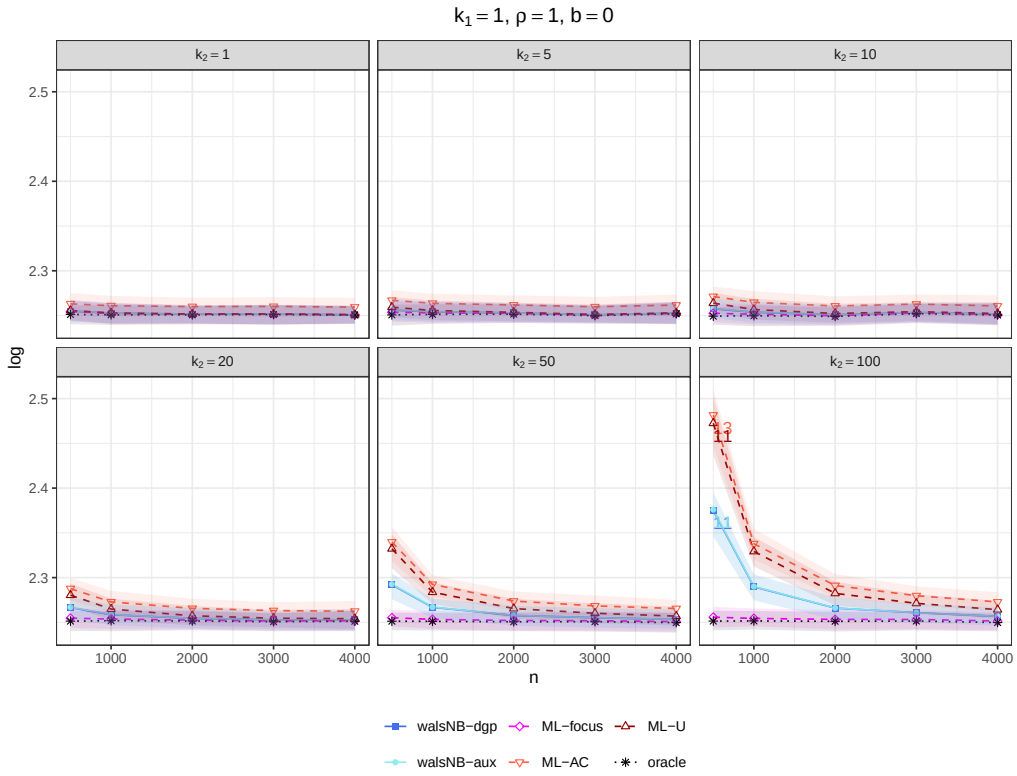

Figure 4: Mean validation log score and quartiles varying n and k_2
The remaining parameters are fixed at $k_1 = 1, \rho = 1$ and $b = 0$. The shaded areas show the interquartile range. The number below a point indicates how often the method failed to converge in this particular setting.

than ML-U in this scenario as their interquartile ranges do not overlap. For large n the distributional fit of all models is similar because the mean log scores converge to the log score of the oracle.

Secondly, similar to the results for RMSE in Figure 3, I find a small upwards shift of the mean validation log scores in Figure 5 when increasing k_1 given $k_2 = 1$. As expected, the distributional fit of ML-AC, which only includes the auxiliary regressors, is the worst among the procedures when $k_1 \geq 5$. Finally, the interquartile range of all models except ML-AC overlap, so their performance in terms of log score typically does not differ. The relative ranking of the procedures for the Brier and spherical score is the same as for the log score, so their results are only shown in the supplementary materials.

In summary, the WALS NB specifications generally outperform ML-U in terms of RMSE and log score when the number of auxiliary regressors is very large relative to the number of focus regressors and when the number of observations is small. This is in line with the results from Abadie and Kasy (2019) for the pretest estimator, which is the predecessor of the WALS estimator: The authors consider a 'Spike and Normal' process for noisy estimates $\hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2, \ldots, \hat{\mathcal{X}}_k$ of e.g. the coefficients from a linear regression model (Abadie and Kasy, 2019, p. 746): The estimates $\hat{\mathcal{X}}_j$ are assumed to follow $\hat{\mathcal{X}}_j \sim \mathcal{N}(m_j, s_j^2)$ for $j = 1, 2, \ldots, k$, e.g. $\hat{\mathcal{X}}_j$ are elements of the ordinary least squares (OLS) estimator in a linear regression model with homoskedastic normal error terms. In this setup, the mean m_j can be regarded as the true value of the regression coefficient that is estimated as $\hat{\mathcal{X}}_j$. The idea is that regularized

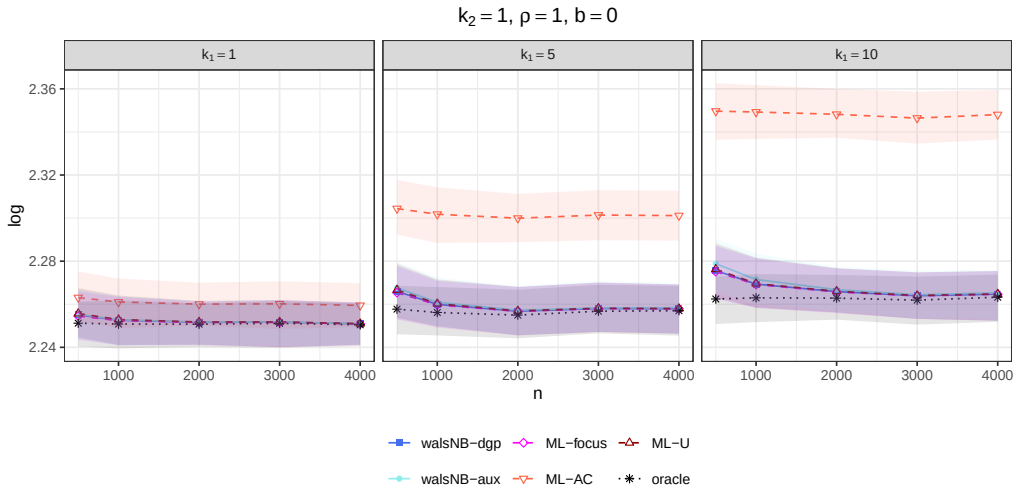

Figure 5: Mean validation log score and quartiles varying n and k_1
The remaining parameters are fixed at $k_2 = 1, \rho = 1$ and $b = 0$. The shaded areas show the interquartile range.

estimators such as lasso, ridge, and pretest modify the OLS estimator $\hat{\mathcal{X}}_j$. Furthermore, the mean m_j is set to zero (spike) with a fixed probability p, and with probability $1 - p$ the coefficient follows $m_j \sim \mathcal{N}(m_0, s_0^2)$ for all j. Under this setting, the authors show that the pretest estimator exhibits smaller integrated risk (integrated expected squared error over the space of distributions of the data distribution, see Abadie and Kasy (2019, p. 745 f.) for details) than lasso and ridge, when the process is very sparse, i.e. p is high and m_0 is large, so many coefficients are set to zero and the non-zero coefficients are far away from zero. The results further agree with the Monte Carlo simulations of De Luca et al. (2023) for WALS in the linear regression model: The authors find that the ratio of the MSE of OLS relative to the MSE of WALS increases when the number of auxiliary regressors k_2 becomes larger. Moreover, for all k_2, the ratio decreases when the sample size n increases. Both observations are in line with Figure 2, where walsNB-dgp dominates in terms of (R)MSE compared to the unrestricted estimator ML-U when n is small and k_2 is large.

6.2 Varying ρ and b

I fixed $k_1 = k_2 = 5$ so that varying the correlation b affects the correlation within focus and auxiliary regressors, as well as the correlation between focus and auxiliary regressors. In contrast, if I had set $k_1 = k_2 = 1$, only the correlation between focus and auxiliary regressors would be modified.

Figure 6 shows that for fixed $b = 0$ and $k_1 = k_2 = 5$, all procedures yield similar mean validation RMSE across all ρ, except for ML-AC, which exhibits much higher values compared to the other methods. Note that the mean validation RMSE generally decreases for all procedures, even the oracle, when ρ increases. This is due to the fact that higher ρ leads to less overdispersion, i.e. lower conditional variance, resulting in lower RMSE for all procedures.

Increasing the correlation b between all regressors for fixed $\rho = 1$ and $k_1 = k_2 = 5$ in Figure 7, the mean validation RMSE shifts down for all procedures, especially for ML-AC

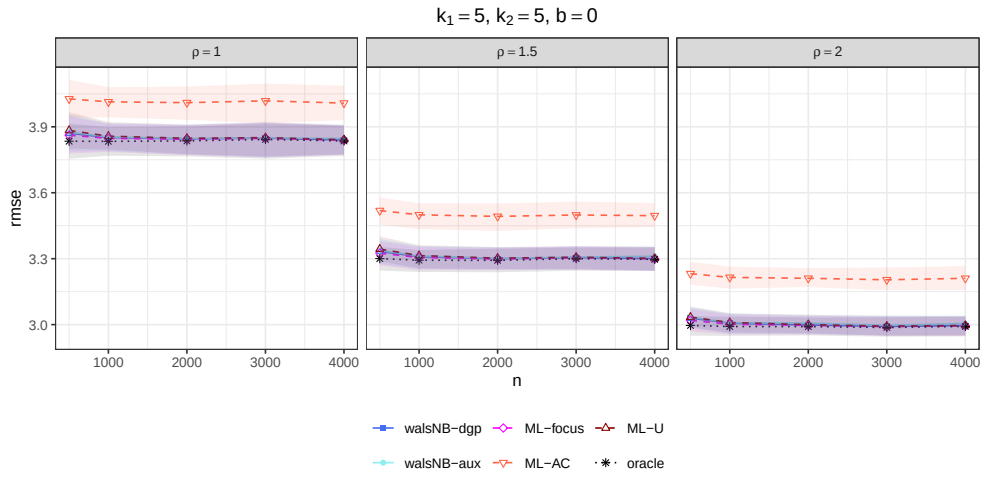

Figure 6: Mean validation RMSE and quartiles varying n and ρ
The remaining parameters are fixed at $b = 0$ and $k_1 = k_2 = 5$. The shaded areas show the interquartile range.

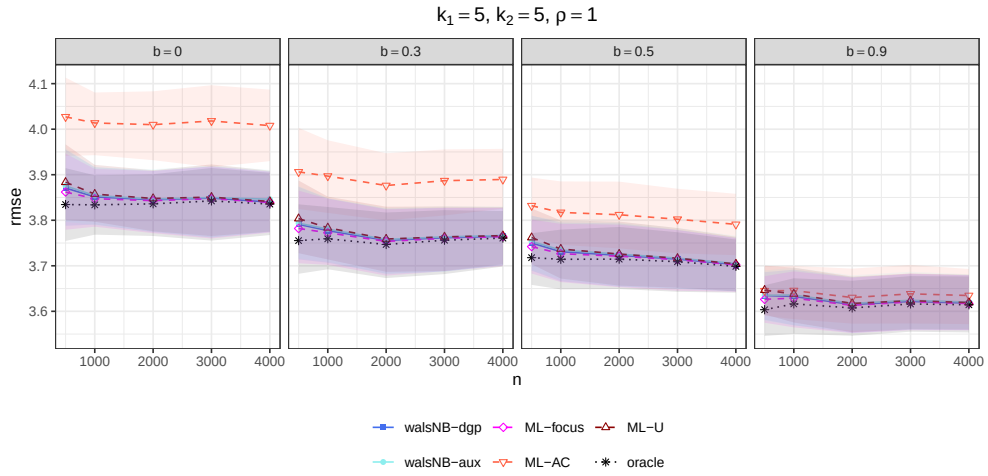

Figure 7: Mean validation RMSE and quartiles varying n and b
The remaining parameters are fixed at $\rho = 1$ and $k_1 = k_2 = 5$. The shaded areas show the interquartile range.

as it only includes the auxiliary regressors and a constant. The larger the correlation, the better it can compensate the lack of focus regressors. Generally, the choice of regressors matters less for prediction when the regressors are highly correlated as each of them will contain similar information for the prediction task. The remaining procedures perform very similarly when increasing b and converge to the mean validation RMSE of the oracle for large n. Except for ML-AC in the cases with $b < 0.9$, the typical RMSE of the procedures are comparable as the interquartile ranges overlap and have similar widths in all scenarios.

The results for the mean validation log score when varying ρ and b with fixed $k_1 = k_2 = 5$ in Figure 8 and Figure 9 are qualitatively the same as for the mean validation RMSE. Interestingly, I also observe a downward shift in the mean validation log score for all procedures and n when I increase ρ. The argument used to explain the downward shift for the mean validation RMSE, namely that the variance around the conditional mean is lower the higher ρ, does not hold anymore since less overdispersion does not necessarily

lead to lower log scores. Intuitively, less overdispersion leads to less variation around the conditional mean that could allow for a more precise estimation of the conditional mean, resulting in an improved distributional fit and, hence, a lower log score.

Finally, Figure 9 shows the mean validation log scores varying n and the correlation b. Similar to the RMSE, the mean validation log scores generally decrease across all n, when b is increased. The reduction is especially large for ML-AC due to the same reasons as for the RMSE in Figure 7. The remaining procedures perform very similarly: Their mean validation log scores are similar and converge to the oracle when n is large and their interquartile ranges overlap.

The relative ranking of the procedures for the Brier and spherical score is similar to that for the log score, so their plots are only shown in the supplementary materials.

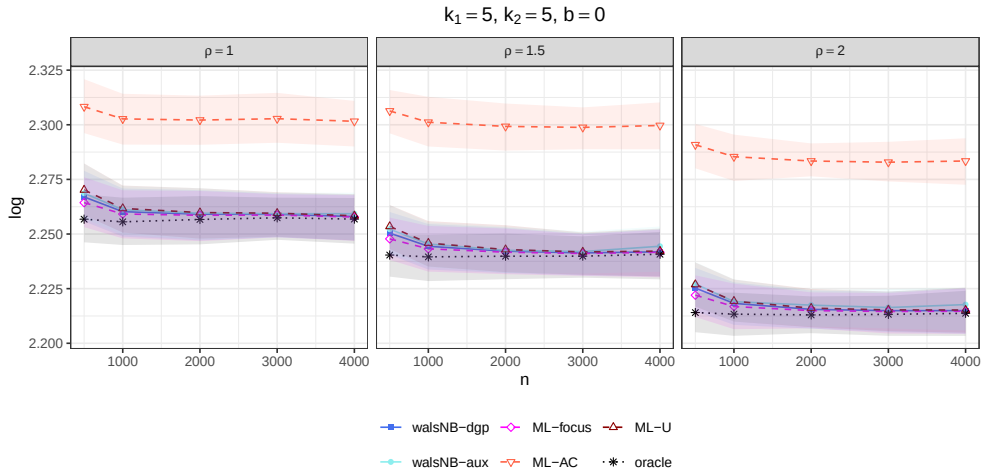

Figure 8: Mean validation log score and quartiles varying n and ρ
The remaining parameters are fixed at $b = 0$ and $k_1 = k_2 = 5$. The shaded areas show the interquartile range.

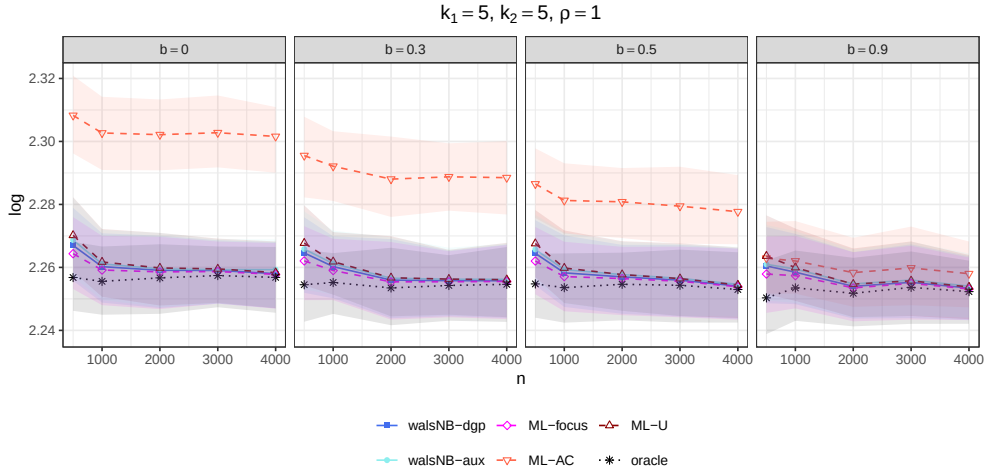

Figure 9: Mean validation log score and quartiles varying n and b
The remaining parameters are fixed at $\rho = 1$ and $k_1 = k_2 = 5$. The shaded areas show the interquartile range.

7 Empirical illustration

The aim of the empirical illustration is to compare the predictive performance of WALS NB with ML and lasso estimation of the NB2 regression model on real data, and to check whether the observations from the simulation experiment translate to a real-world application.

I use the cross-sectional data set called 'DoctorVisits', which derives from the 1977-1978 Australian Health Survey and was analyzed in Cameron and Trivedi (1986) and Mullahy (1997). The dataset contains $n = 5190$ observations from individuals over 18 years of age on twelve variables, including the response `visits`, which describes the number of doctor visits in a two-week period before the interview. It further provides explanatory variables such as income and age, as well as health-related variables like recent illnesses and health insurance coverage. The data are available via the R package **AER** (Kleiber and Zeileis, 2008) as `DoctorVisits` based on the original from the Journal of Applied Econometrics Data Archive.[6]

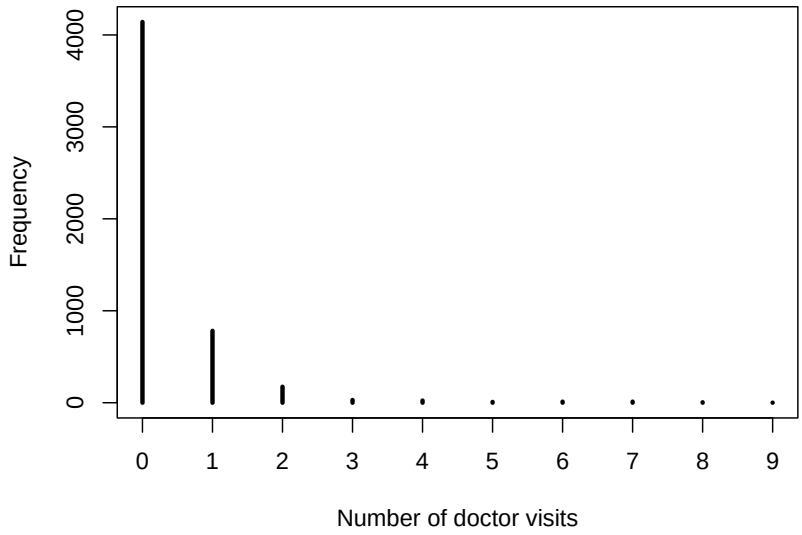

Figure 10: Visualization of `visits` in DoctorVisits

Table D.1 and D.2 in Appendix D provide a description and summary statistics of the variables in the DoctorVisits dataset. Further, Figure 10 shows a visualization of the response `visits`, which clearly exhibits overdispersion and will be modeled using regression models for count data. For the computation of the Brier and spherical score, I truncate the infinite sum in $||\hat{p}||$ from (5.4) at the largest observed count of the dataset, which is 9.

Inspired by the applications of Rupp et al. (2012, p. 2 f.) and Faber et al. (2020, p. 164 f.) in quantum chemistry, I apply K-fold cross-validation (CV) to produce 'learning curves' that allow me to compare the performance of the procedures for different sizes of the training set. Algorithm 1 illustrates the process for generating a K-fold cross-validated learning curve for any evaluation metric and procedure.

[6] `https://www.journaldata.zbw.eu/dataset/heterogeneity-excess-zeros-and-the-structure-of-count-data-models`

Algorithm 1 K-fold cross-validated learning curves

1. Randomly split dataset $\mathcal{D} := \{(y_i, x_i)\}_{i=1,2,\dots,n}$ into K parts $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K$ of roughly the same size (see implementation of `cv()` of **mboost** (Hofner et al., 2014) for more details, size of last partition will be smaller if n/K is not an integer). Then, the training and validation set $\mathcal{T}_k$ and $\mathcal{V}_k$ for each fold $k = 1, 2, \dots, K$ are defined as

$$\mathcal{T}_k := \{\mathcal{D}_j : j \neq k\}, \quad \mathcal{V}_k := \mathcal{D}_k.$$

   Further, let $\tau_k : \{1, 2, \dots, |\mathcal{T}_k|\} \to \{1, 2, \dots, n\}$ be an indexing function that maps the index of the observations of $\mathcal{T}_k$ to the original dataset $\mathcal{D}$.

2. Specify the grid for the number of training observations $t = (t_1, t_2, \dots, t_L)$ with $t_L \leq t_{max}$, where $t_{max} = |\mathcal{D}| - \max_k |\mathcal{D}_k|$ is the largest possible size of the training set.

3. For $l = 1, 2, \dots, L$:

   (a) For procedure $m = 1, 2, \dots, M$:

      i. For $k = 1, 2, \dots, K$:

         A. Fit and tune procedure m on data $\mathcal{T}_{l,k} := \{(y_{\tau_k(r)}, x_{\tau_k(r)})\}_{r=1,2,\dots,t_l}$.

         B. Compute validation metric $\hat{s}_{l,m,k}$ on $\mathcal{V}_k$.

      ii. Output the cross-validated metric for training size t_l: $\hat{s}_{l,m} = \frac{1}{K} \sum_{k=1}^{K} \hat{s}_{l,m,k}$.

4. The learning curve for each $m = 1, 2, \dots, M$ plots $\hat{s}_{l,m}$ against t_l for $l = 1, 2, \dots, L$.

Note that only the training sets $\mathcal{T}_{l,k}$ vary in size but the validation sets $\mathcal{V}_k$ remain the same. Similar to Meek et al. (2002, p. 398), the training sets $\mathcal{T}_{l,k}$ are nested, i.e. $\mathcal{T}_{l,k} \subset \mathcal{T}_{l+1,k}$ for $l = 1, 2 \dots, L-1$. For all experiments below, I use $K = 10$ folds.

I compare procedures that differ in the estimator and specification of the mean, where the choices for the latter are inspired by the applications in Cameron and Trivedi (1986, p. 46 ff.). The different combinations of estimator and specification are named following the pattern: 'estimator-specification'. Again, 'walsNB' and 'ML' represent the WALS NB and ML estimator, respectively, while 'lasso' estimates the NB2 regression model using the lasso estimator of Wang et al. (2016) (see Appendix B for details). I consider a total of six estimator-specification combinations:

1. walsNB-main: Includes all covariates linearly as auxiliary regressors and only a constant as focus regressor.

2. walsNB-main-focus: Includes all covariates and a constant linearly as focus regressors, a quadratic term for age and two-way interactions between health and gender, health and age, health and income, and finally gender and illness as auxiliary regressors.

3. walsNB-int: Includes all regressors of walsNB-main-focus (including interactions) as auxiliary and only a constant as focus regressor.

4. ML-main: Includes all covariates linearly and a constant and hence uses the same regressors as walsNB-main.

5. ML-int: Includes all regressors of ML-main but adds a quadratic term for age and two-way interactions between health and gender, health and age, health and income, and finally gender and illness. Counterpart of walsNB-main-focus and walsNB-int.

6. lasso-int: Includes all covariates linearly, a quadratic term for age and two-way interactions between health and gender, health and age, health and income, and finally gender and illness *in the fitting process*. Depending on the choice of the regularization parameter, not all the aforementioned regressors have to be included in the final model. In contrast, a constant is always included.

All procedures are fitted using a log-link for the mean parameter. WALS NB procedures use the Laplace prior because the Weibull led to numerical instabilities in some small subsamples resulting from the numerical integration procedure required for computing the posterior mean of the auxiliary regression coefficients in (4.4). The parameters of the Laplace prior are taken from Magnus and De Luca (2016, p. 132), which are minimax regret solutions for the normal location problem. The remaining settings for WALS NB and ML specifications are retained from the simulation experiment of Section 6. Notably, all WALS NB specifications use the unrestricted ML estimator for NB2 as starting values for the regression coefficients and the dispersion parameter. By unrestricted, I refer to the unrestricted model given the covariates that are included in the specification, i.e. ML-main for walsNB-main and ML-int for, both, walsNB-main-focus and walsNB-int.

The lasso specification 'lasso-int' performs tuning (maximizing 10-fold CV log-likelihood) in the training set $\mathcal{T}_{l,k}$ of each fold k (and each training set size t_l), as recommended by Hothorn et al. (2005, p. 679) who include tuning and final model fit in the estimation process. This is sensible, as tuning of the regularization parameter is key to the performance of lasso. Different values of the regularization parameter correspond to different levels of regularization and the regressors included in the model may also differ. Moreover, the method also standardizes the regressors in the training set of each fold to have zero mean and unit variance before estimation (i.e., it uses the estimated means and variances of the regressors in the subsample and not over the entire dataset).

In Figure 11 and Table 2 we observe that all WALS NB specifications except for walsNB-main-focus outperform the ML specifications in terms of RMSE for all numbers of training observations. The differences are particularly large for small training sets, e.g. for $t_l = 500$ the CV RMSE of walsNB-int is almost 19% smaller than ML-int. Except for $t_l < 1500$, walsNB-int and walsNB-main also outperform the lasso specification lasso-int. Further, note that walsNB-int outperforms walsNB-main-focus although the only difference between the two procedures is that the latter specifies some of the covariates as focus regressors. This observation seems to contradict the results of the simulation experiment, where walsNB-dgp and walsNB-aux perform very similarly even though the latter considers all covariates as auxiliary regressors and the former considers part of them as focus regressors. However, walsNB-dgp chooses the same focus regressors as the DGP of the simulation, which is unlikely in empirical applications.

The CV log scores in Figure 12 and Table 3 show that all procedures perform similarly in terms of the distributional fit. Moreover, the curves decrease as I increase the number of

Table 2: 10-fold CV RMSE varying t_l, DoctorVisits

Training obs. t_l	500	1000	1500	2000	2500	3000	3500	4000	4671
walsNB-main	0.838	0.886	0.807	0.756	0.747	0.746	0.750	0.756	0.760
walsNB-main-focus	1.012	1.068	0.893	0.789	0.768	0.765	0.767	0.774	0.780
walsNB-int	0.837	0.864	0.797	0.753	0.745	0.742	0.743	0.746	0.752
ML-main	1.013	0.950	0.873	0.794	0.771	0.766	0.774	0.781	0.785
ML-int	1.032	0.956	0.885	0.792	0.769	0.762	0.764	0.770	0.776
lasso-int	0.791	0.826	0.831	0.772	0.761	0.753	0.759	0.764	0.776

– All figures rounded to three decimal places.

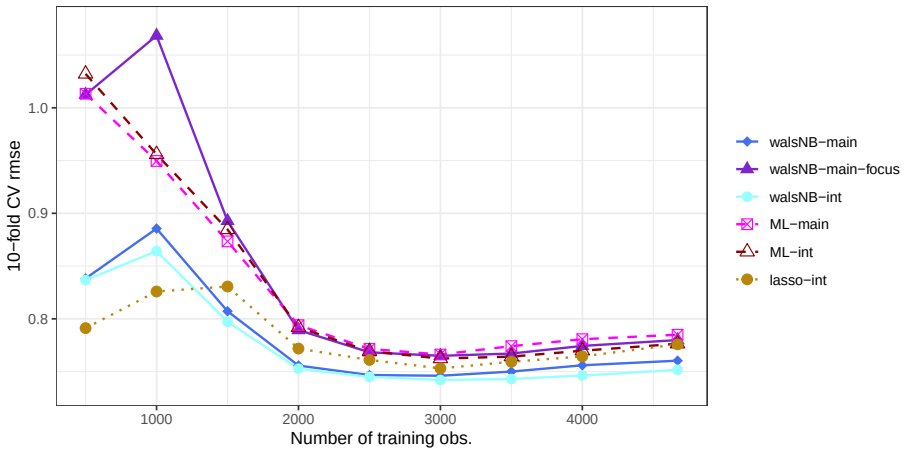

Figure 11: 10-fold CV RMSE varying t_l, DoctorVisits

training observations and flatten at about 2000 observations. This shows that the methods are able to 'learn' more (i.e. improve the fit in terms of log score), when more training observations are available but stop 'learning' at some point, i.e. when the curves flatten.

The other metrics for distributional fit, Brier and spherical score, show qualitatively similar results but the curves are flatter, hence the results are only shown in the supplementary materials. This further underlines that the distributional fit of the methods does not improve drastically when the dataset becomes larger.

Note that WALS NB specifications are computationally less demanding than lasso, while performing similarly in terms of CV RMSE and log score. They do not require any tuning unlike lasso, which performs an 'internal' 10-fold CV to choose the optimal regularization parameter. Consequently, the fitting times of WALS NB are typically shorter than those of lasso and competitive with the ML specifications. Of course, one may change the parameters of the fitting algorithm of lasso to improve the computing time. However, it should not result in better performance metrics as the current setup already favors lasso: It allows many iterations in the fitting algorithm and, thus, a high chance for convergence.

In conclusion, all WALS NB specifications, except for walsNB-main-focus, perform better than the ML specifications in terms of RMSE, while metrics for the distributional fit such as log, Brier and spherical score are similar or minimally worse than the ML specifications. Moreover, the RMSE is similar or slightly lower than for lasso at large t_l, while demanding less computational resources as WALS NB does not require tuning by CV.

Table 3: 10-fold CV log score varying t_l, DoctorVisits

Training obs. t_l	500	1000	1500	2000	2500	3000	3500	4000	4671
walsNB-main	0.639	0.628	0.623	0.622	0.622	0.622	0.621	0.621	0.620
walsNB-main-focus	0.642	0.629	0.624	0.622	0.621	0.621	0.621	0.620	0.620
walsNB-int	0.642	0.629	0.626	0.625	0.623	0.623	0.623	0.622	0.621
ML-main	0.640	0.627	0.622	0.621	0.620	0.621	0.620	0.620	0.620
ML-int	0.644	0.630	0.626	0.624	0.622	0.622	0.622	0.621	0.621
lasso-int	0.637	0.626	0.623	0.622	0.622	0.622	0.621	0.620	0.620

– All figures rounded to three decimal places.

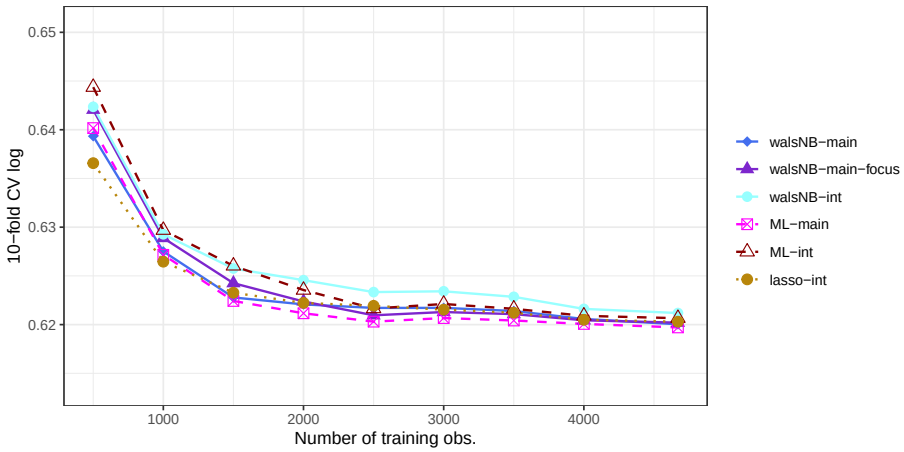

Figure 12: 10-fold CV log score varying t_l, DoctorVisits

8 Conclusion

This paper extends the WALS approach to NB2 regression models (WALS NB) for count data based on WALS GLM of De Luca et al. (2018) and compares its predictive performance to the traditional ML and lasso estimator in simulated and real count datasets using the classical measure RMSE and strictly proper scoring rules.

In the simulation experiment, WALS NB outperforms the ML estimator in very sparse situations, i.e. where the number of auxiliary regressors is large and the number of training observations is small. When increasing the number of training observations, WALS NB and the unrestricted ML estimator converge in all performance metrics. Interestingly, whether WALS NB includes all regressors as auxiliary regressors or parts of them as focus does not change the results substantially. This shows that specifying all regressors as auxiliary is a reasonable choice if no prior information is available on the importance of the individual regressors. Moreover, it highlights that the regularized Bayesian estimation of the coefficients of the auxiliary regressors is key for the performance of WALS NB.

The empirical illustration emphasizes the results found in the simulation experiment: For small training sets, WALS NB using all covariates as auxiliary regressors outperforms all ML specifications in terms of RMSE while yielding a comparable distributional fit measured by strictly proper scores. Only the lasso estimator yields lower RMSE for small training sets. However, it is more computationally demanding than WALS NB due to the

additional 10-fold CV that is run for determining the optimal regularization parameter. This highlights an important advantage of WALS compared to other model averaging techniques: low computational costs. Moreover, WALS NB using all covariates as auxiliary regressors outperformed all other specifications of WALS NB. Thus, if only the predictive power is of concern, WALS NB is a viable alternative to established estimation methods for the NB2 regression model that is easy to specify (choose all regressors as auxiliary), regularized and computationally efficient.

For future research, it would be interesting to generalize WALS to hurdle or zero-inflation models to handle count data with excess zeros. Thus far, WALS has been limited to univariate response variables, therefore extending the methodology to multivariate outcomes would allow a larger variety of applications, such as joint modeling of related count processes. Lastly, an investigation of the large sample properties of WALS could improve our understanding of statistical inference after model averaging.

Acknowledgements

The scientific computing center sciCORE (`https://scicore.unibas.ch/`) at the University of Basel provided me with valuable computing resources. I would also like to thank Christian Kleiber for our helpful discussions.

Appendix

A Proofs

Proof of Proposition 3.1. I start by rewriting the equation system from (3.5) using the data transformations in (3.6). Notice that

$$\bar{X}_p^\top \bar{V}(y-\bar{\mu}) = \sum_{i=1}^{n} \bar{v}_i(y_i - \bar{\mu}_i)x_{ip},$$
$$\bar{X}_p^\top \bar{X}_q = \sum_{i=1}^{n} \bar{\psi}_i x_{ip} x_{iq}^\top,$$

for $p, q = 1, 2$, so the first equation of (3.5) can be expressed as

$$\begin{aligned} 0 = \bar{X}_1^\top \Big(&\bar{X}_1\bar{\beta}_1 + \bar{X}_2\bar{\beta}_2 + \bar{\Psi}^{-1/2}\bar{V}(y-\bar{\mu}) - \bar{g}\bar{\Psi}^{-1/2}\bar{C}(y-\bar{\mu})\bar{\alpha} - \bar{X}_1\beta_1 - \bar{X}_2\beta_2 \\ &+ \bar{g}\bar{\Psi}^{-1/2}\bar{C}(y-\bar{\mu})\alpha \Big). \end{aligned}$$

Using $\bar{y}_0$ from (3.6), the expression can be written more compactly as

$$0 = \bar{X}_1^\top \left(\bar{y}_0 - \bar{X}_1\beta_1 - \bar{X}_2\beta_2 + \bar{g}\bar{\Psi}^{-1/2}\bar{C}(y-\bar{\mu})\alpha \right). \tag{A.1}$$

Following the same steps, the second equation of (3.5) becomes

$$0 = \bar{X}_2^\top \left(\bar{y}_0 - \bar{X}_1\beta_1 - \bar{X}_2\beta_2 + \bar{g}\bar{\Psi}^{-1/2}\bar{C}(y-\bar{\mu})\alpha \right) - R_j\nu_j. \tag{A.2}$$

Analogously, the third equation in (3.5) can be expressed as

$$\begin{aligned} 0 &= \bar{g}\bar{\kappa}^\top \mathbf{1} - \bar{g}(y-\bar{\mu})^\top \bar{C}\bar{\eta} + \bar{g}(y-\bar{\mu})^\top \bar{C}(X_1\beta_1 + X_2\beta_2) + (\bar{g}^2\bar{k}^\top \mathbf{1} + \bar{\varrho}\bar{\kappa}^\top \mathbf{1})(\alpha - \bar{\alpha}), \\ &= \bar{t} + \bar{g}(y-\bar{\mu})^\top \bar{C}(X_1\beta_1 + X_2\beta_2) + (\bar{g}^2\bar{k}^\top \mathbf{1} + \bar{\varrho}\bar{\kappa}^\top \mathbf{1})\alpha. \end{aligned} \tag{A.3}$$

Assuming $\bar{g}^2\bar{k}^\top \mathbf{1} + \bar{\varrho}\bar{\kappa}^\top \mathbf{1} \neq 0$, I can solve (A.3) for α:

$$\alpha = -\frac{\bar{t} + \bar{g}(y-\bar{\mu})^\top \bar{C}(X_1\beta_1 + X_2\beta_2)}{\bar{g}^2\bar{k}^\top \mathbf{1} + \bar{\varrho}\bar{\kappa}^\top \mathbf{1}}. \tag{A.4}$$

Let us combine (A.1) and (A.2) into a larger matrix equation. First, move some terms so they become

$$\begin{aligned} \bar{X}_1^\top \bar{X}_1\beta_1 + \bar{X}_1^\top \bar{X}_2\beta_2 &= \bar{X}_1^\top \bar{y}_0 + \bar{X}_1^\top \bar{\Psi}^{-1/2}\bar{C}(y-\bar{\mu})\bar{g}\alpha, \\ \bar{X}_2^\top \bar{X}_1\beta_1 + \bar{X}_2^\top \bar{X}_2\beta_2 &= \bar{X}_2^\top \bar{y}_0 + \bar{X}_2^\top \bar{\Psi}^{-1/2}\bar{C}(y-\bar{\mu})\bar{g}\alpha - R_j\nu_j. \end{aligned}$$

Using $\bar{\Psi}^{-1/2}\bar{X}_p = X_p, p = 1, 2$, collect both equations to

$$\begin{pmatrix} \bar{X}_1^\top \bar{X}_1 & \bar{X}_1^\top \bar{X}_2 \\ \bar{X}_2^\top \bar{X}_1 & \bar{X}_2^\top \bar{X}_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \bar{X}_1^\top \bar{y}_0 \\ \bar{X}_2^\top \bar{y}_0 \end{pmatrix} + \begin{pmatrix} X_1^\top \bar{C}(y-\bar{\mu}) \\ X_2^\top \bar{C}(y-\bar{\mu}) \end{pmatrix} \bar{g}\alpha - \begin{pmatrix} 0 \\ R_j \end{pmatrix} \nu_j.$$

Inserting (A.4) and rearranging yields

$$\begin{aligned}\begin{pmatrix}\bar{X}_1^\top\bar{X}_1 & \bar{X}_1^\top\bar{X}_2\\ \bar{X}_2^\top\bar{X}_1 & \bar{X}_2^\top\bar{X}_2\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = &\begin{pmatrix}\bar{X}_1^\top\bar{y}_0\\ \bar{X}_2^\top\bar{y}_0\end{pmatrix} - \frac{\bar{t}\bar{g}}{\bar{g}^2\bar{k}^\top\mathbf{1}+\bar{\varrho}\bar{\kappa}^\top\mathbf{1}}\begin{pmatrix}X_1^\top\bar{C}(y-\bar{\mu})\\ X_2^\top\bar{C}(y-\bar{\mu})\end{pmatrix}\\ &- \frac{\bar{g}^2}{\bar{g}^2\bar{k}^\top\mathbf{1}+\bar{\varrho}\bar{\kappa}^\top\mathbf{1}}\begin{pmatrix}X_1^\top\bar{C}(y-\bar{\mu})(y-\bar{\mu})^\top\bar{C}X_1\beta_1\\ X_2^\top\bar{C}(y-\bar{\mu})(y-\bar{\mu})^\top\bar{C}X_1\beta_1\end{pmatrix}\\ &- \frac{\bar{g}^2}{\bar{g}^2\bar{k}^\top\mathbf{1}+\bar{\varrho}\bar{\kappa}^\top\mathbf{1}}\begin{pmatrix}X_1^\top\bar{C}(y-\bar{\mu})(y-\bar{\mu})^\top\bar{C}X_2\beta_2\\ X_2^\top\bar{C}(y-\bar{\mu})(y-\bar{\mu})^\top\bar{C}X_2\beta_2\end{pmatrix}\\ &- \begin{pmatrix}0\\ R_j\end{pmatrix}\nu_j,\end{aligned}$$

which can be further rewritten, using $\bar{\epsilon}$ and $\bar{q}$ as defined in Section 3.1, to

$$\bar{A}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\bar{X}_1^\top\bar{y}_0\\ \bar{X}_2^\top\bar{y}_0\end{pmatrix} - \bar{t}\bar{\epsilon}\begin{pmatrix}X_1^\top\bar{q}\\ X_2^\top\bar{q}\end{pmatrix} - \begin{pmatrix}0\\ R_j\end{pmatrix}\nu_j, \tag{A.5}$$

with

$$\bar{A} := \begin{pmatrix}\bar{X}_1^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1 & \bar{X}_1^\top\bar{X}_2 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_2\\ \bar{X}_2^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_2^\top\bar{q}\bar{q}^\top X_1 & \bar{X}_2^\top\bar{X}_2 + \bar{g}\bar{\epsilon}X_2^\top\bar{q}\bar{q}^\top X_2\end{pmatrix}.$$

Then, consider the partitioned inverse

$$\bar{A}^{-1} = \begin{pmatrix}\bar{A}^{11} & \bar{A}^{12}\\ \bar{A}^{21} & \bar{A}^{22}\end{pmatrix},$$

with elements

$$\begin{aligned}\bar{A}^{11} = &(\bar{X}_1^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1)^{-1}\\ &+ \Big[(\bar{X}_1^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1)^{-1}(\bar{X}_1^\top\bar{X}_2 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_2)(\bar{X}_2^\top\bar{M}_1\bar{X}_2)^{-1}\\ &\cdot(\bar{X}_2^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_2^\top\bar{q}\bar{q}^\top X_1)(\bar{X}_1^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1)^{-1}\Big],\end{aligned} \tag{A.6}$$

$$\bar{A}^{12} = -(\bar{X}_1^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1)^{-1}(\bar{X}_1^\top\bar{X}_2 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_2)(\bar{X}_2^\top\bar{M}_1\bar{X}_2)^{-1} = \bar{A}^{21\top}, \tag{A.7}$$

$$\bar{A}^{22} = (\bar{X}_2^\top\bar{M}_1\bar{X}_2)^{-1}. \tag{A.8}$$

It is assumed that $\bar{X}_2^\top\bar{M}_1\bar{X}_2$ is positive definite so all elements of the partitioned inverse of $\bar{A}^{-1}$ exist.

Using the Sherman-Morrison-Woodbury formula, I can rewrite the following inverse if $(1+\bar{g}\bar{\epsilon}\bar{q}^\top X_1(\bar{X}_1^\top\bar{X}_1)^{-1}X_1^\top\bar{q}) \neq 0$ as

$$(\bar{X}_1^\top\bar{X}_1 + \bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1)^{-1} = (\bar{X}_1^\top\bar{X}_1)^{-1} - \frac{(\bar{X}_1^\top\bar{X}_1)^{-1}(\bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1)(\bar{X}_1^\top\bar{X}_1)^{-1}}{1+\bar{g}\bar{\epsilon}\bar{q}^\top X_1(\bar{X}_1^\top\bar{X}_1)^{-1}X_1^\top\bar{q}}. \tag{A.9}$$

Therefore, $(\bar{X}_1^\top\bar{X}_1+\bar{g}\bar{\epsilon}X_1^\top\bar{q}\bar{q}^\top X_1)^{-1}$ exists if $(1+\bar{g}\bar{\epsilon}\bar{q}^\top X_1(\bar{X}_1^\top\bar{X}_1)^{-1}X_1^\top\bar{q}) \neq 0$ and $\bar{X}_1^\top\bar{X}_1$ is invertible. The latter is easily shown because $\text{rank}(\bar{X}_1) = \text{rank}(\bar{\Psi}^{-1/2}X_1) = \text{rank}(X_1) = k_1$ (assumed X_1 to have full column rank) and $\text{rank}(\bar{\Psi}^{-1/2}) = n$, otherwise I would not be

able to compute $\bar{\Psi}^{-1/2}$. Thus, $\bar{X}_1$ has full column rank and $\text{rank}(\bar{X}_1^\top \bar{X}_1) = k_1$.

Let $\tilde{\beta}_{1u}$ and $\tilde{\beta}_{2u}$ denote the solution of the unrestricted model, then plugging $R_u = 0$ into (A.5) yields the unrestricted equation system

$$\bar{A}\begin{pmatrix}\tilde{\beta}_{1u}\\ \tilde{\beta}_{2u}\end{pmatrix} = \begin{pmatrix}\bar{X}_1^\top \bar{y}_0\\ \bar{X}_2^\top \bar{y}_0\end{pmatrix} - \bar{t}\bar{\epsilon}\begin{pmatrix}X_1^\top \bar{q}\\ X_2^\top \bar{q}\end{pmatrix}. \tag{A.10}$$

Further, let $\tilde{\beta}_{1j}$ and $\tilde{\beta}_{2j}$ denote the solution of the jth model. Using (A.5) then yields

$$\bar{A}\begin{pmatrix}\tilde{\beta}_{1j}\\ \tilde{\beta}_{2j}\end{pmatrix} = \begin{pmatrix}\bar{X}_1^\top \bar{y}_0\\ \bar{X}_2^\top \bar{y}_0\end{pmatrix} - \bar{t}\bar{\epsilon}\begin{pmatrix}X_1^\top \bar{q}\\ X_2^\top \bar{q}\end{pmatrix} - \begin{pmatrix}0\\ R_j\end{pmatrix}\tilde{\nu}_j. \tag{A.11}$$

Combining (A.10) and (A.11), I can find an explicit solution for ν_j as they imply

$$\bar{A}\begin{pmatrix}\tilde{\beta}_{1j}\\ \tilde{\beta}_{2j}\end{pmatrix} = \bar{A}\begin{pmatrix}\tilde{\beta}_{1u}\\ \tilde{\beta}_{2u}\end{pmatrix} - \begin{pmatrix}0\\ R_j\end{pmatrix}\tilde{\nu}_j,$$

then multiply with $\bar{A}^{-1}$ so

$$\begin{pmatrix}\tilde{\beta}_{1j}\\ \tilde{\beta}_{2j}\end{pmatrix} = \begin{pmatrix}\tilde{\beta}_{1u}\\ \tilde{\beta}_{2u}\end{pmatrix} - \begin{pmatrix}\bar{A}^{11} & \bar{A}^{12}\\ \bar{A}^{21} & \bar{A}^{22}\end{pmatrix}\begin{pmatrix}0\\ R_j\end{pmatrix}\tilde{\nu}_j. \tag{A.12}$$

Multiply both sides by $\begin{pmatrix}0 & R_j^\top\end{pmatrix}$ and note that by (3.3) $R_j^\top \tilde{\beta}_{2j} = 0$, then

$$\begin{aligned}\begin{pmatrix}0 & R_j^\top\end{pmatrix}\begin{pmatrix}\tilde{\beta}_{1j}\\ \tilde{\beta}_{2j}\end{pmatrix} &= \begin{pmatrix}0 & R_j^\top\end{pmatrix}\begin{pmatrix}\tilde{\beta}_{1u}\\ \tilde{\beta}_{2u}\end{pmatrix} - \begin{pmatrix}0 & R_j^\top\end{pmatrix}\begin{pmatrix}\bar{A}^{11} & \bar{A}^{12}\\ \bar{A}^{21} & \bar{A}^{22}\end{pmatrix}\begin{pmatrix}0\\ R_j\end{pmatrix}\tilde{\nu}_j\\ 0 &= R_j^\top \tilde{\beta}_{2u} - R_j^\top \bar{A}^{22} R_j \tilde{\nu}_j\\ \rightarrow \tilde{\nu}_j &= (R_j^\top \bar{A}^{22} R_j)^{-1} R_j^\top \tilde{\beta}_{2u},\end{aligned} \tag{A.13}$$

assuming the inverse $(R_j^\top \bar{A}^{22} R_j)^{-1}$ exists. Plug (A.13) into (A.12) for

$$\begin{aligned}\tilde{\beta}_{1j} &= \tilde{\beta}_{1u} - \bar{A}^{12} R_j (R_j^\top \bar{A}^{22} R_j)^{-1} R_j^\top \tilde{\beta}_{2u},\\ \tilde{\beta}_{2j} &= \tilde{\beta}_{2u} - \bar{A}^{22} R_j (R_j^\top \bar{A}^{22} R_j)^{-1} R_j^\top \tilde{\beta}_{2u}.\end{aligned}$$

Now, insert (A.7), (A.8) and introduce n so

$$\begin{aligned}\tilde{\beta}_{1j} = \tilde{\beta}_{1u} + \Bigg[&\left(\frac{\bar{X}_1^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_1\right)^{-1}\left(\frac{\bar{X}_1^\top \bar{X}_2}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_2\right)\left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n}\right)^{-1/2}\\ &\cdot \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n}\right)^{-1/2} R_j \left(R_j^\top \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n}\right)^{-1} R_j\right)^{-1} R_j^\top \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n}\right)^{-1/2}\\ &\cdot \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n}\right)^{1/2} \tilde{\beta}_{2u}\Bigg]\\ = \tilde{\beta}_{1u} + \bar{Q}\bar{P}_j\tilde{\vartheta}.\end{aligned} \tag{A.14}$$

Moreover, introduce n also for $\tilde{\beta}_{2j}$ which yields

$$\begin{aligned}\tilde{\beta}_{2j} &= \tilde{\beta}_{2u} - \Bigg[\left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1} R_j \left(R_j^\top \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1} R_j \right)^{-1} R_j^\top \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2} \\ &\quad \cdot \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{1/2} \tilde{\beta}_{2u} \Bigg] \\ &= \tilde{\beta}_{2u} - \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2} \bar{P}_j \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{1/2} \tilde{\beta}_{2u} \\ &= \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2} (I_{k_2} - \bar{P}_j) \tilde{\vartheta} \\ &= \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2} \bar{W}_j \tilde{\vartheta}.\end{aligned}$$

In order to obtain the estimator of the fully restricted model $\tilde{\beta}_{1r}$, I set $R_r = I_{k_2}$. Combined with (A.14) this leads to

$$\tilde{\beta}_{1r} = \tilde{\beta}_{1u} + \bar{Q}\tilde{\vartheta} \tag{A.15}$$

$$\leftrightarrow \tilde{\beta}_{1u} = \tilde{\beta}_{1r} - \bar{Q}\tilde{\vartheta}. \tag{A.16}$$

Insert (A.16) in (A.14) to get

$$\tilde{\beta}_{1j} = \bar{\beta}_{1r} - \bar{Q}\bar{W}_j\tilde{\vartheta}.$$

What remains to be derived are $\tilde{\beta}_{1u}$ and $\tilde{\beta}_{2u}$. First, multiply (A.10) with $\bar{A}^{-1}$ for

$$\begin{pmatrix} \tilde{\beta}_{1u} \\ \tilde{\beta}_{2u} \end{pmatrix} = \begin{pmatrix} \bar{A}^{11} & \bar{A}^{12} \\ \bar{A}^{21} & \bar{A}^{22} \end{pmatrix} \begin{pmatrix} \bar{X}_1^\top \bar{y}_0 \\ \bar{X}_2^\top \bar{y}_0 \end{pmatrix} - \bar{t}\bar{\epsilon} \begin{pmatrix} \bar{A}^{11} & \bar{A}^{12} \\ \bar{A}^{21} & \bar{A}^{22} \end{pmatrix} \begin{pmatrix} X_1^\top \bar{q} \\ X_2^\top \bar{q} \end{pmatrix}.$$

Inserting (A.6), (A.7) and (A.8) results in

$$\begin{aligned}\tilde{\beta}_{1u} = \Bigg[&(\bar{X}_1^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_1)^{-1} \\ &+ \Bigg\{ (\bar{X}_1^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_1)^{-1} (\bar{X}_1^\top \bar{X}_2 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_2)(\bar{X}_2^\top \bar{M}_1 \bar{X}_2^\top)^{-1} \\ &(\bar{X}_2^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_2^\top \bar{q}\bar{q}^\top X_1)(\bar{X}_1^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_1)^{-1} \Bigg\} \Bigg] (\bar{X}_1^\top \bar{y}_0 - \bar{t}\bar{\epsilon} X_1^\top \bar{q}) \\ &- (\bar{X}_1^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_1)^{-1} (\bar{X}_1^\top \bar{X}_2 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_2)(\bar{X}_2^\top \bar{M}_1 \bar{X}_2^\top)^{-1} (\bar{X}_2^\top \bar{y}_0 - \bar{t}\bar{\epsilon} X_2^\top \bar{q}).\end{aligned} \tag{A.17}$$

and

$$\begin{aligned}\tilde{\beta}_{2u} = &- (\bar{X}_2^\top \bar{M}_1 \bar{X}_2)^{-1} (\bar{X}_2^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_2^\top \bar{q}\bar{q}^\top X_1)(\bar{X}_1^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_1)^{-1} (\bar{X}_1^\top \bar{y}_0 - \bar{t}\bar{\epsilon} X_1^\top \bar{q}) \\ &+ (\bar{X}_2^\top \bar{M}_1 \bar{X}_2)^{-1} (\bar{X}_2^\top \bar{y}_0 - \bar{t}\bar{\epsilon} X_2^\top \bar{q}).\end{aligned} \tag{A.18}$$

Plugging $\tilde{\beta}_{1u}$ and $\tilde{\beta}_{2u}$ into (A.15) and exploiting (3.7) yields

$$\begin{aligned}\tilde{\beta}_{1r} &= (\bar{X}_1^\top \bar{X}_1 + \bar{g}\bar{\epsilon} X_1^\top \bar{q}\bar{q}^\top X_1)^{-1}(\bar{X}_1^\top \bar{y}_0 - \bar{t}\bar{\epsilon} X_1^\top \bar{q}) \\ &= \left(\frac{\bar{X}_1^\top \bar{X}_1}{n} + \frac{\bar{g}\bar{\epsilon}}{n} X_1^\top \bar{q}\bar{q}^\top X_1\right)^{-1} \left(\frac{\bar{X}_1^\top \bar{y}_0}{n} - \frac{\bar{t}\bar{\epsilon}}{n} X_1^\top \bar{q}\right).\end{aligned} \tag{A.19}$$

Finally, insert $\tilde{\beta}_{1j}$ and $\tilde{\beta}_{2j}$ in (A.4) for the estimator of the dispersion parameter

$$\tilde{\alpha}_j = -\frac{\bar{t} + \bar{g}(y - \bar{\mu})^\top \bar{C}(X_1\tilde{\beta}_{1j} + X_2\tilde{\beta}_{2j})}{\bar{g}^2 \bar{k}^\top \mathbf{1} + \bar{\varrho}\bar{\kappa}^\top \mathbf{1}}.$$

The proposition collects the solutions for $\tilde{\beta}_{1j}$, $\tilde{\beta}_{2j}$, $\tilde{\alpha}_j$ and $\tilde{\beta}_{1r}$. □

Proof of Corollary 3.2. I prove the first statement in Corollary 3.2 by contradiction. Assume for $j \neq \{u, r\}$ that the jth model for the transformed regressors Z and untransformed regressors X are equivalent so we can transform the estimators for the auxiliary variables into each other, i.e. $\tilde{\gamma}_{2j} = \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\tilde{\beta}_{2j}$. The assumption requires the restrictions in the estimation for the transformed and untransformed regressors (see (3.5)) to be equivalent, i.e.

$$R_j^\top \tilde{\gamma}_{2j} = R_j^\top \tilde{\beta}_{2j}. \tag{$*$}$$

Without loss of generality, I analyze the special case $k_2 = 2$, where $\tilde{\gamma}_{2j} = (\tilde{\gamma}_{2j,1}, \tilde{\gamma}_{2j,2})^\top$ and $\tilde{\beta}_{2j} = (\tilde{\beta}_{2j,1}, \tilde{\beta}_{2j,2})^\top$, and the jth model is assumed to set $\tilde{\gamma}_{2j,2} = \tilde{\beta}_{2j,2} = 0$ via the restriction matrix $R_j^\top = (0, 1)$. Define the elements of $\bar{\Xi}^{1/2}$ to be

$$\bar{\Xi}^{1/2} = \begin{pmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{pmatrix},$$

and $\bar{\Delta}_2^{-1}$ reduces to a 2×2 diagonal matrix $\bar{\Delta}_2^{-1} = \text{diag}(\bar{\Delta}_2^{11}, \bar{\Delta}_2^{22})$.

Under the assumption $\tilde{\gamma}_{2j} = \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\tilde{\beta}_{2j}$, the restriction in the estimation of the jth model for the transformed regressors is

$$\begin{aligned}R_j^\top \tilde{\gamma}_{2j} &= R_j^\top \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\tilde{\beta}_{2j} \\ &= \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{pmatrix} \begin{pmatrix} \bar{\Delta}_2^{11}\tilde{\beta}_{2j,1} \\ \bar{\Delta}_2^{22}\tilde{\beta}_{2j,2} \end{pmatrix} \\ &= \xi_{21}\bar{\Delta}_2^{11}\tilde{\beta}_{2j,1} + \xi_{22}\bar{\Delta}_2^{22}\tilde{\beta}_{2j,2} = 0,\end{aligned}$$

while the restriction in the estimation of the jth model for the untransformed regressors is

$$R_j^\top \tilde{\beta}_{2j} = \tilde{\beta}_{2j,2} = 0.$$

Therefore, generally $R_j^\top \tilde{\gamma}_{2j} \neq R_j^\top \tilde{\beta}_{2j}$ so the restrictions in the estimator using the transformed and untransformed regressors differ, which contradicts ($*$). Consequently, $\tilde{\gamma}_{2j} = \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\tilde{\beta}_{2j}$ cannot generally hold for $k_2 > 2$ as it does not even hold for $k_2 = 2$. By extension, $\tilde{\gamma}_{1j} \neq \bar{\Delta}_1^{-1}\tilde{\beta}_{1j}$ because the restrictions in the estimation differ, which finishes the

proof of the first part of the corollary.

For $k_2 = 1$, there exist only two models: 1. the unrestricted and 2. the fully restricted model so $j \in \{u, r\}$. In this case, the general results $\tilde{\gamma}_{1u} = \bar{\Delta}^{-1}\tilde{\beta}_{1u}$, $\tilde{\gamma}_{2u} = \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\tilde{\beta}_{2u}$ and $\tilde{\gamma}_{1r} = \bar{\Delta}_1^{-1}\tilde{\beta}_{1r}$ from (3.13) apply. Finally, the fully restricted model fulfills $R_r^\top \tilde{\gamma}_{2r} = \tilde{\gamma}_{2r} = 0$, which finishes the proof of the second part of the corollary. □

B Software and implementation

The simulation experiment of Section 6 is performed on the scientific computing center sciCORE at the University of Basel (`https://scicore.unibas.ch/`) with R version 4.3.0 (R Core Team, 2023), while the empirical illustration of Section 7 is computed on a local machine running R version 4.3.1. Models estimated by WALS are fitted using the newly developed R package **WALS** version 0.2.4 (Huynh, 2023) available from the Comprehensive R Archive Network (CRAN, `https://cran.r-project.org/package=WALS`). **WALS** is based on the MATLAB code version 2.0 for WALS in the linear regression model by Magnus and De Luca (2016) which can be downloaded from `https://www.janmagnus.nl/items/WALS.pdf`. The dependencies of **WALS** along with the particular versions used are: **Formula** version 1.2-5 (Zeileis and Croissant, 2010), **MASS** version 7.3-60 (Venables and Ripley, 2002) and **Rdpack** version 2.5 (Boshnakov, 2023). Standard NB2 regressions use `glm.nb()` from **MASS**. The function uses an algorithm that alternates between fitting the coefficients β of the NB2 regression for fixed dispersion parameter ρ using IRLS (iteratively reweighted least squares) and then maximizing the log-likelihood with respect to ρ given β. Moreover, training and validation splits for K-fold CV in Section 7 are generated using the function `cv()` of **mboost** version 2.9-8 (Hofner et al., 2014) and the computations are parallelized over the number of training observations t_l using **parallel**. The lasso estimation of 'lasso-int' in Section 7 is performed using `cv.glmregNB()` from **mpath** version 0.4-2.23 (Wang, 2023).

Finally, the integral in (4.4) is evaluated numerically for all priors except for the Laplace prior. Numerical integration is performed using the `integrate()` function of **stats**, which uses an adaptive quadrature method with the basic step being a Gauss-Kronrod quadrature (for more details see the code documentation).

Computational details

The lasso estimator of the NB2 regression model maximizes the following penalized objective function from Wang et al. (2016, p. 2687 f.):

$$\max_{\beta,\rho} L(\beta, \rho) = \max_{\beta,\rho} \left\{ \ell(\beta, \rho) - n \cdot d \sum_{j=1}^{p} |\beta_j| \right\},$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, β_0 is the constant and $d \geq 0$ is the regularization parameter. Moreover, $\ell(\beta, \rho)$ is the log-likelihood from (3.1) and the penalty term is scaled by the sample size n such that the penalty does not vanish when the number of observations becomes large. Notice that neither the constant β_0 nor the dispersion parameter ρ are regularized, but only the 'true' regression coefficients β_j, $j > 0$, are regularized.

The optimal regularization parameter d for 'lasso-int' in Section 7 is determined by maximizing the 10-fold CV log-likelihood, where candidates $\{d_1, d_2, \ldots, d_H\}$ are generated by first running the method on the entire training set (ignoring the folds) and setting d_H such that an intercept only model is estimated. Then, the minimum value is set as $d_1 = a \cdot d_H$, $0 < a < 1$. The remaining values are determined by an evenly spaced grid on the log scale, i.e. between $\log(d_1)$ and $\log(d_H)$.

The implementation alternates between P iterations of 1. the coordinate descent algorithm described in Wang et al. (2016, p. 2690 f.) to estimate the regression coefficients β given a value of the dispersion parameter ρ and 2. maximizing the log-likelihood with respect to ρ given the estimate of β. The latter uses `theta.ml()` from **MASS** (Venables and Ripley, 2002) but limits its number of iterations to 10. The alternation process is stopped when certain convergence criteria are met. For the estimation of 'lasso-int' in Section 7, the maximum allowed number of iterations (until convergence) for the coordinate descent algorithm is increased to 2500 from the default setting of 1000. Furthermore, I also increase the maximum number of alternations between coordinate descent and ML estimation of the dispersion parameter ρ from 10 to 1000. The remaining settings are left at their default values, see the documentation of **mpath** (Wang, 2023) for more details.

A caveat for the results of 'lasso-int' on the DoctorVisits dataset is that the ML estimation of ρ often reaches its internal iteration limit within the alternation process between coordinate descent, for estimating the regression coefficients, and ML estimation of ρ using `theta.ml()` from **MASS**. Ideally, I could increase the number of iterations used in `theta.ml()`, however, the implementation of lasso in **mpath** uses a fixed number of 10 iterations. I compensate for this by allowing the alternation process to run a maximum of 1000 iterations (instead of the default 10) until convergence before it is forced to terminate. The idea is that, even if `theta.ml()` does not converge in an iteration of the alternation process, the alternation process can go through many iterations, where `theta.ml()` is run in each round.

Tables are generated using **xtable** version 1.8-4 (Dahl et al., 2019) and **stargazer** version 5.2.3 (Hlavac, 2022), and some plots use **ggplot2** version 3.4.4 (Wickham, 2016) with themes from **ggthemes** version 4.2.4 (Arnold, 2021). LaTeX expressions are inserted with **latex2exp** version 0.9.6 (Meschiari, 2022). Finally, results are partly processed with **abind** version 1.4-5 (Plate and Heiberger, 2016) before plotting.

C Additional tables for the simulation experiment

Table C.1: Values of $\bar{\beta}_1$

1	2	3	4	5	6	7	8	9	10
-0.1518	-0.197	0.1401	0.1328	-0.155	0.1775	0.1403	0.1272	0.1778	-0.1602

– All figures rounded to four decimal places.

Table C.2: Values of $\bar{\beta}_2$

	1	2	3	4	5	6	7	8	9	10
0	-8×10^{-4}	0.0089	0.0052	0.0087	-6×10^{-4}	0.0021	-3×10^{-4}	-0.0078	-0.005	0
10	-0.0025	0.0087	5×10^{-4}	-0.0037	-0.0044	0.0058	0.004	-0.0067	-0.0087	0.0051
20	0.0024	-0.0066	-0.0088	-0.0078	-0.0024	-0.0066	-0.004	-0.0062	-0.0049	-0.0064
30	-5×10^{-4}	0.0054	-0.0094	5×10^{-4}	0.0076	-0.0025	-0.009	-0.0072	-0.0036	-0.0069
40	-0.0074	-0.0056	-0.0055	-0.0074	0.0096	-0.0035	1×10^{-4}	0.0036	-0.008	-0.0076
50	-0.009	0.0086	0.0035	-0.0081	-1×10^{-4}	-8×10^{-4}	-0.0025	0.0098	-0.0065	0.0063
60	-0.0086	-0.002	-0.0072	-0.0061	0.0068	0.0044	-0.0047	-1×10^{-4}	-0.0083	-0.0029
70	0.0094	0.0025	0.0033	-0.0038	-0.0019	0.0099	0.0071	0.0091	0.0062	0.0056
80	-0.0046	0.0052	0.0097	-0.0041	-0.002	0.0062	-0.0085	-0.0027	-0.0011	-0.0069
90	0.0016	0.0094	0.0098	-0.0065	8×10^{-4}	-0.0023	0.0035	-0.0046	-6×10^{-4}	-0.0066

– To get element $\bar{\beta}_{2,z}$, construct z as the sum of the row and the column.
– Example: $\bar{\beta}_{2,23} = -0.0088$.
– All figures rounded to four decimal places.

D Additional tables for the empirical illustration

Table D.1: Variable descriptions for DoctorVisits

Variable	Description
visits	# of doctor visits in past two weeks.
genderfemale	= 1 if individual is female. Omitted reference category: male.
age	Age in years divided by 100.
income	Annual income in tens of thousands of dollars.
illness	# of illnesses in past two weeks.
reduced	# of days of reduced activity in past two weeks due to illness or injury.
health	General health questionnaire score using Goldberg's method.
privateyes	= 1 if individual has private health insurance Omitted reference category: individual has no private health insurance.
freepooryes	= 1 if individual has free government health insurance due to low income.
freerepatyes	= 1 if individual has free government health insurance due to old age, disability or veteran status. Omitted reference category: individual has no free government health insurance.
nchronicyes	= 1 if individual has a chronic condition which does not limit activity.
lchronicyes	= 1 if individual has a chronic condition which limits activity. Omitted reference category: individual has no chronic condition.

[–] Reproduced based on the documentation of `DoctorVisits` in **AER** (Kleiber and Zeileis, 2008).

Table D.2: Summary statistics for DoctorVisits

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
visits	0.302	0.798	0	0	0	0	9
health	1.218	2.124	0	0	0	2	12
genderfemale	0.521	0.500	0	0	1	1	1
age	0.406	0.205	0.190	0.220	0.320	0.620	0.720
income	0.583	0.369	0	0.250	0.550	0.900	1.500
illness	1.432	1.384	0	0	1	2	5
reduced	0.862	2.888	0	0	0	0	14
privateyes	0.443	0.497	0	0	0	1	1
freepooryes	0.043	0.202	0	0	0	0	1
freerepatyes	0.210	0.407	0	0	0	0	1
nchronicyes	0.403	0.491	0	0	0	1	1
lchronicyes	0.117	0.321	0	0	0	0	1
age^2	0.207	0.186	0.036	0.048	0.102	0.384	0.518
health×genderfemale	0.695	1.752	0	0	0	0	12
health×age	0.503	1.013	0	0	0	0.570	8.640
health×income	0.643	1.378	0	0	0	0.750	16.500
genderfemale×illness	0.839	1.312	0	0	0	1	5

– St. Dev. : Standard deviation, Pctl(25): 25% quantile, Pctl(75): 75% quantile.
– × indicates interaction between two variables.
– Number of observations $N = 5190$.
– All figures rounded to three decimal places.

Supplementary materials

S1 Asymptotic distribution of one-step ML estimators

Following De Luca et al. (2018, p. 4), I use the local misspecification framework with true auxiliary parameters set to

$$\beta_2 = \frac{\delta}{\sqrt{n}}, \tag{S1.1}$$

so the true parameter vector $\varphi = (\beta_1^\top, \beta_2^\top, \alpha)^\top$ converges to $\varphi_0 = (\beta_1, 0, \alpha)$ for $n \to \infty$. Note that the asymptotics in this paper all refer to the case of $n \to \infty$.

I start the derivation by noting the asymptotic distribution of the fully iterated ML estimator of the unrestricted model $\check{\beta}_{1u}$, $\check{\beta}_{2u}$ and $\check{\alpha}_u$. Under the usual ML regularity conditions (see e.g. Crowder, 1976), it holds

$$\sqrt{n}\begin{pmatrix} \check{\beta}_{1u} - \beta_1 \\ \check{\beta}_{2u} - \beta_2 \\ \check{\alpha}_u - \alpha \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Omega), \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} & \Omega_{1\alpha} \\ \Omega_{21} & \Omega_{22} & \Omega_{2\alpha} \\ \Omega_{\alpha 1} & \Omega_{\alpha 2} & \Omega_{\alpha\alpha} \end{pmatrix} := \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} & \mathcal{I}_{1\alpha} \\ \mathcal{I}_{21} & \mathcal{I}_{22} & \mathcal{I}_{2\alpha} \\ \mathcal{I}_{\alpha 1} & \mathcal{I}_{\alpha 2} & \mathcal{I}_{\alpha\alpha} \end{pmatrix}^{-1} = \mathcal{I}^{-1},$$

where $\mathcal{I} := \mathcal{I}(\varphi_0)$ is the information matrix evaluated at φ_0. For the remainder, I will restrict the analysis to the one-step ML estimators for the regression coefficients. The asymptotic distribution including the dispersion coefficient can be derived analogously.

Let $\bar{\varphi} = (\bar{\beta}_1^\top, \bar{\beta}_2^\top, \bar{\alpha})^\top$ collect the starting values and assume $\bar{\varphi} - \varphi = O_p(1/\sqrt{n})$, then the unrestricted one-step ML estimator has the same asymptotic distribution as the fully iterated ML estimator (see e.g. Theorem 3.5 in Newey and McFadden, 1994), i.e.

$$\sqrt{n}\begin{pmatrix} \tilde{\beta}_{1u} - \beta_1 \\ \tilde{\beta}_{2u} - \beta_2 \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Omega_S), \quad \Omega_S = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}. \tag{S1.2}$$

Using $\mathcal{I}_{p\alpha} = 0 = \mathcal{I}_{\alpha p}^\top$, $p = 1, 2$ (Lawless, 1987, p. 211), the elements of Ω_S can be expressed as

$$\begin{aligned} \Omega_{11} &= \mathcal{I}_{11}^{-1} + \mathcal{I}_{11}^{-1}\mathcal{I}_{12}(\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12})^{-1}\mathcal{I}_{21}\mathcal{I}_{11}^{-1}, \\ \Omega_{12} &= -\mathcal{I}_{11}^{-1}\mathcal{I}_{12}(\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12})^{-1} = \Omega_{21}^\top, \\ \Omega_{22} &= (\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12})^{-1}. \end{aligned} \tag{S1.3}$$

If the DGP is included in the set of models considered for averaging, using the fully iterated ML estimator of the unrestricted model as starting values ensures, under the usual ML regularity conditions, that $\bar{\varphi} - \varphi = O_p(1/\sqrt{n})$ (De Luca et al., 2018, p. 4).

The following proposition utilizes the aforementioned ingredients and provides the asymptotic distribution of the one-step ML estimators for the jth model:

Proposition S1.1 (Asymptotic distribution of one-step ML estimators)**.** *In addition to* (S1.1)*, assume that the usual ML regularity conditions hold (see e.g. Crowder, 1976). If* $\bar{\varphi} - \varphi = O_p(1/\sqrt{n})$*, then as* $n \to \infty$*,*

$$\sqrt{n}\begin{pmatrix}\tilde{\beta}_{1j} - \beta_1 \\ \tilde{\beta}_{2j} - \beta_2\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}\mathcal{Q}\mathcal{P}_j\Omega_{22}^{-1/2}\delta \\ -\Omega_{22}^{1/2}\mathcal{P}_j\Omega_{22}^{-1/2}\delta\end{pmatrix}, \begin{pmatrix}\mathcal{I}_{11}^{-1} + \mathcal{Q}\mathcal{W}_j\mathcal{Q}^\top & -\mathcal{Q}\mathcal{W}_j\Omega_{22}^{1/2} \\ -\Omega_{22}^{1/2}\mathcal{W}_j\mathcal{Q}^\top & \Omega_{22}^{1/2}\mathcal{W}_j\Omega_{22}^{1/2}\end{pmatrix}\right),$$

where $\mathcal{I}_{pq}$ *denotes the pqth submatrix of* $\mathcal{I}(\varphi_0)$*. Further,* $\Omega_{22} = (\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12})^{-1}$*,* $\mathcal{Q} = \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\Omega_{22}^{1/2}$*,* $\mathcal{P}_j = \Omega_{22}^{1/2}R_j(R_j^\top\Omega_{22}R_j)^{-1}R_j^\top\Omega_{22}^{1/2}$ *and* $\mathcal{W}_j = I_{k_2} - \mathcal{P}_j$*.*

Proof of Proposition S1.1. Let me first consider the asymptotic distribution of $\tilde{\vartheta}$. Equation (3.7) implies

$$\sqrt{n}(\tilde{\vartheta} - \vartheta) = \left(\frac{\bar{X}_2^\top\bar{M}_1\bar{X}_2}{n}\right)^{1/2}\sqrt{n}(\tilde{\beta}_{2u} - \beta_2) + \left[\left(\frac{\bar{X}_2^\top\bar{M}_1\bar{X}_2}{n}\right)^{1/2} - \Omega_{22}^{-1/2}\right]\delta,$$

where $\vartheta := \Omega_{22}^{-1/2}\beta_2$. Next, I analyze the probability limit of $\bar{X}_2^\top\bar{M}_1\bar{X}_2/n$. As $n \to \infty$

$$\begin{aligned}\operatorname{plim}\frac{\bar{X}_2^\top\bar{M}_1\bar{X}_2}{n} = \operatorname{plim}\Bigg(&\frac{\bar{H}_{22}}{n} - \frac{1}{n}\frac{\bar{H}_{2\alpha}\bar{H}_{\alpha 2}}{\bar{H}_{\alpha\alpha}} \\ &- \left\{\left(\frac{\bar{H}_{21}}{n} - \frac{1}{n}\frac{\bar{H}_{2\alpha}\bar{H}_{\alpha 1}}{\bar{H}_{\alpha\alpha}}\right)\left(\frac{\bar{H}_{11}}{n} - \frac{1}{n}\frac{\bar{H}_{1\alpha}\bar{H}_{\alpha 1}}{\bar{H}_{\alpha\alpha}}\right)^{-1}\right. \\ &\left.\cdot\left(\frac{\bar{H}_{12}}{n} - \frac{1}{n}\frac{\bar{H}_{1\alpha}\bar{H}_{\alpha 2}}{\bar{H}_{\alpha\alpha}}\right)\right\}\Bigg) \\ = \mathcal{I}_{22} &- \frac{\mathcal{I}_{2\alpha}\mathcal{I}_{\alpha 2}}{\mathcal{I}_{\alpha\alpha}} \\ &- \left(\mathcal{I}_{21} - \frac{\mathcal{I}_{2\alpha}\mathcal{I}_{\alpha 1}}{\mathcal{I}_{\alpha\alpha}}\right)\left(\mathcal{I}_{11} - \frac{\mathcal{I}_{1\alpha}\mathcal{I}_{\alpha 1}}{\mathcal{I}_{\alpha\alpha}}\right)^{-1}\left(\mathcal{I}_{12} - \frac{\mathcal{I}_{1\alpha}\mathcal{I}_{\alpha 2}}{\mathcal{I}_{\alpha\alpha}}\right).\end{aligned}$$

Lawless (1987, p. 211) derived in equation (2.7b), that $\mathcal{I}_{p\alpha} = 0$, $p = 1, 2$. Combined with (S1.3) this yields

$$\operatorname{plim}\frac{\bar{X}_2^\top\bar{M}_1\bar{X}_2}{n} = \Omega_{22}^{-1}. \tag{S1.4}$$

Thus, using (S1.2) I get

$$\sqrt{n}(\tilde{\vartheta} - \vartheta) \xrightarrow{d} \mathcal{N}(0, I_{k_2}). \tag{S1.5}$$

Now, I show the asymptotic distribution for all restricted estimators. First, I need the asymptotic distribution of the fully restricted estimator. From (A.15) I deduce

$$\sqrt{n}(\tilde{\beta}_{1r} - \beta_1) = \sqrt{n}(\tilde{\beta}_{1u} - \beta_1) + \bar{Q}\sqrt{n}(\tilde{\vartheta} - \vartheta) + \bar{Q}\Omega_{22}^{-1/2}\delta.$$

Since

$$\operatorname{plim}\bar{Q} = \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\Omega_{22}^{1/2} =: \mathcal{Q},$$

which is the same as in Proposition 2 of De Luca et al. (2018, p. 4) for WALS GLM because

α is asympotically independent of β_1 and β_2 in the NB2 model, it follows that

$$\sqrt{n}(\tilde{\beta}_{1r} - \beta_1) \xrightarrow{d} \mathcal{N}(\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\delta, \mathcal{I}_{11}^{-1}). \tag{S1.6}$$

This is equivalent to the asymptotic distribution of the fully restricted one-step ML estimator of the WALS GLM model in equation (A.4) of De Luca et al. (2018, p. 14). Furthermore, $\tilde{\beta}_{1r}$ and $\tilde{\vartheta}$ are also asymptotically independent as their joint asymptotic distribution is a multivariate normal with covariance $\Omega_{12}\Omega_{22}^{-1/2} + \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\Omega_{22}^{1/2} \overset{(S1.3)}{=} 0$.

Finally, I have all ingredients to derive the asymptotic distribution of the general one-step ML estimator. Proposition 3.1 implies

$$\sqrt{n}(\tilde{\beta}_{1j} - \beta_1) = \bar{Q}\bar{P}_j\Omega_{22}^{-1/2}\delta + (\sqrt{n}(\tilde{\beta}_{1r} - \beta_1) - \bar{Q}\Omega_{22}^{-1/2}\delta) - \bar{Q}\bar{W}_j\sqrt{n}(\tilde{\vartheta} - \vartheta),$$

and

$$\sqrt{n}(\tilde{\beta}_{2j} - \beta_2) = \left[\left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n}\right)^{-1/2} \bar{W}_j\Omega_{22}^{-1/2} - I_{k_2}\right]\delta + \left(\frac{\bar{X}_2^\top \bar{M}_1 \bar{X}_2}{n}\right)^{-1/2} \bar{W}_j\sqrt{n}(\tilde{\vartheta} - \vartheta).$$

Together with (S1.5) and (S1.6), the fact that $\tilde{\beta}_{1r}$ and $\tilde{\vartheta}$ are asymptotically independent, and the probability limits

$$\begin{aligned}\operatorname{plim} \bar{P}_j &= \Omega_{22}^{1/2} R_j (R_j^\top \Omega_{22} R_j)^{-1} R_j^\top \Omega_{22}^{1/2} =: \mathcal{P}_j, \\ \operatorname{plim} \bar{W}_j &= I_{k_2} - \mathcal{P}_j =: \mathcal{W}_j,\end{aligned}$$

they imply the joint asymptotic distribution of $\tilde{\beta}_{1j}$ and $\tilde{\beta}_{2j}$ from the proposition. □

S2 Asymptotic distribution of one-step ML estimators in transformed models

In order to establish the asymptotic distribution of the one-step ML estimators in the transformed models, it is necessary to determine the probability limits of the matrices involved in the transformation from X to Z. Firstly, $\bar{\Delta}_1$ converges in probability to a constant matrix Δ_1, i.e.

$$\operatorname{plim} \bar{\Delta}_1 =: \Delta_1,$$

because $\operatorname{plim} \bar{X}_1^\top \bar{X}_1/n = \mathcal{I}_{11}$ is constant and $\bar{\Delta}_1 = \operatorname{diag}\left(\bar{X}_1^\top \bar{X}_1/n\right)^{-1/2}$. By the same line of reasoning,

$$\operatorname{plim} \bar{\Delta}_2 =: \Delta_2,$$

which is constant because $\bar{\Delta}_2 = \operatorname{diag}\left(\bar{X}_2^\top \bar{M}_1 \bar{X}_2/n\right)^{-1/2}$ and $\operatorname{plim} = \bar{X}_2^\top \bar{M}_1 \bar{X}_2/n \overset{(S1.4)}{=} \Omega_{22}^{-1}$ is constant. Thus, the probability limit of $\bar{\Xi}$ follows as

$$\operatorname{plim} \bar{\Xi} = \operatorname{plim} \frac{\bar{\Delta}_2 \bar{X}_2^\top \bar{M}_1 \bar{X}_2 \bar{\Delta}_2}{n} \overset{(S1.4)}{=} \Delta_2 \Omega_{22}^{-1} \Delta_2 =: \Xi.$$

Similar to De Luca et al. (2018, p. 5), let $\gamma_{10} = \Delta^{-1}\beta_1$ and $\gamma_{2n} = \Xi^{1/2}\Delta_2^{-1}\beta_2$, then

Proposition S1.1 implies

$$\sqrt{n}\begin{pmatrix}\tilde{\gamma}_{1j}-\gamma_{10}\\ \tilde{\gamma}_{2j}-\gamma_{2n}\end{pmatrix}\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix}\mathcal{D}P_j d\\ -P_j d\end{pmatrix},\begin{pmatrix}\mathcal{J}_{11}^{-1}+\mathcal{D}W_j\mathcal{D}^\top & -\mathcal{D}W_j\\ -W_j\mathcal{D}^\top & W_j\end{pmatrix}\right), \tag{S2.1}$$

where $d = \Xi^{1/2}\Delta_2^{-1}\delta$, $\mathcal{D} = \operatorname{plim}\bar{D} = \mathcal{J}_{11}^{-1}\mathcal{J}_{12}$, $\mathcal{J}_{11} = \operatorname{plim}\bar{Z}_1^\top\bar{Z}_1/n = \Delta_1\mathcal{I}_{11}\Delta_1$ and $\mathcal{J}_{12} = \operatorname{plim}\bar{Z}_1^\top\bar{Z}_2/n = \Delta_1\mathcal{I}_{12}\Delta_2\Xi^{-1/2}$. Therefore, I can approximate the distribution of $\sqrt{n}\tilde{\gamma}_{2u}$ in large samples with

$$\sqrt{n}\tilde{\gamma}_{2u} \approx \mathcal{N}(\sqrt{n}\gamma_{2n}, I_{k_2}) = \mathcal{N}(d, I_{k_2}).$$

S3 Additional results for the simulation experiment

S3.1 Results for alternative scoring rules

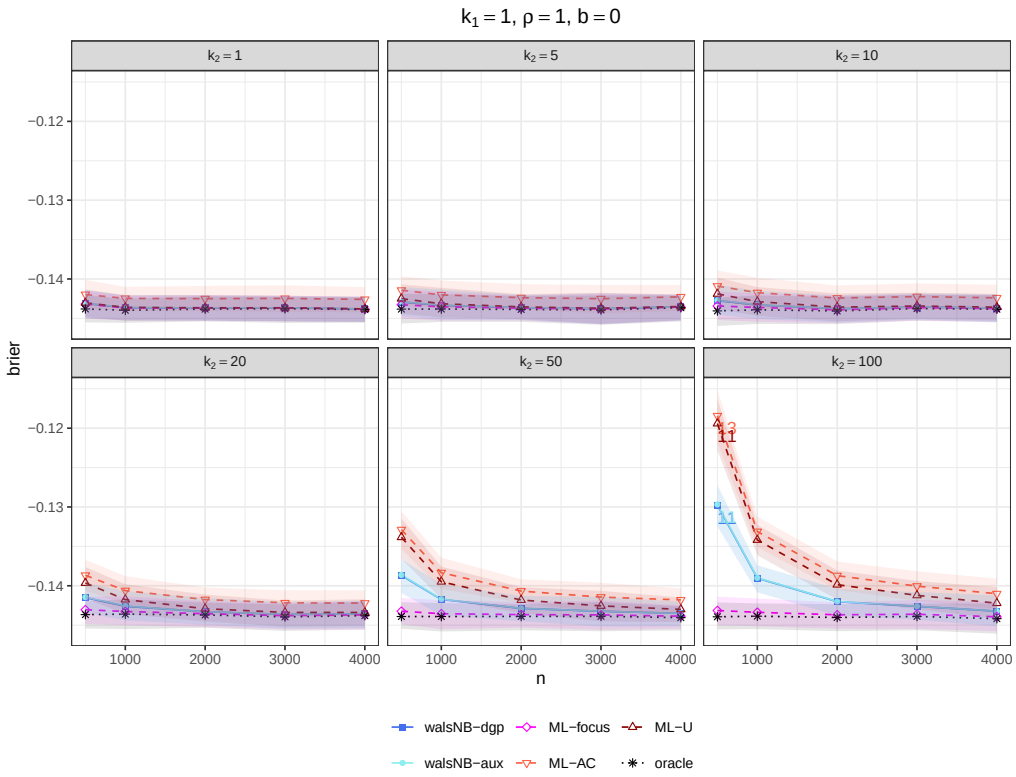

Figure S3.1: Mean validation Brier score and quartiles varying n and k_2
The remaining parameters are fixed at $k_1 = 1, \rho = 1$ and $b = 0$. Each point represents the mean over all successful runs of the experiment, i.e. over $R = 300$ when it never fails to converge. The shaded areas show the interquartile range. The number below a point indicates how often the method failed to converge in this particular setting.

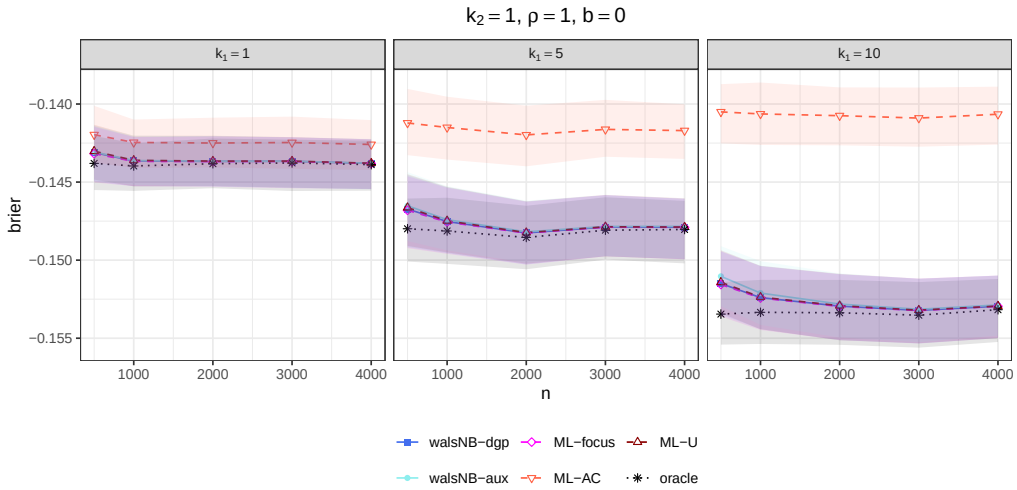

Figure S3.2: Mean validation Brier score and quartiles varying n and k_1
The remaining parameters are fixed at $k_2 = 1, \rho = 1$ and $b = 0$. Each point represents the mean over all successful runs of the experiment, i.e. over $R = 300$ when it never fails to converge. The shaded areas show the interquartile range.

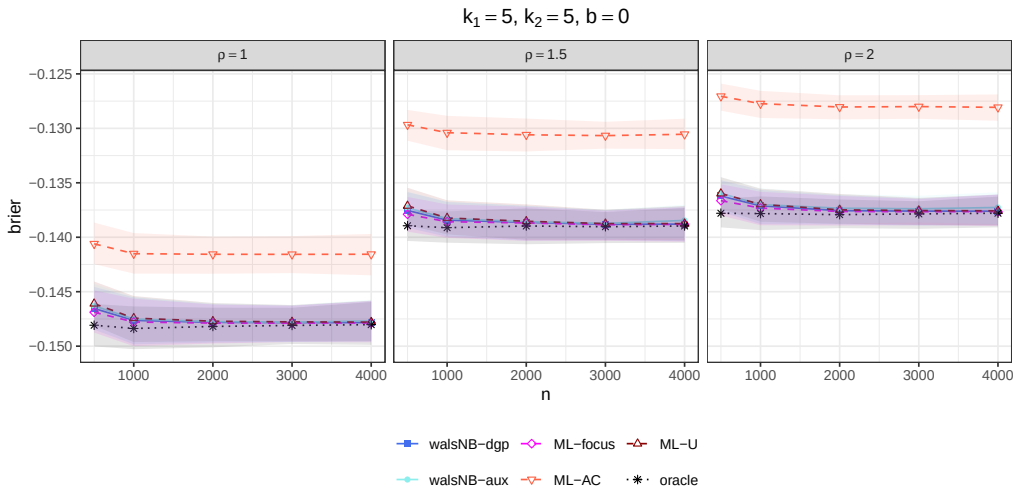

Figure S3.3: Mean validation Brier score and quartiles varying n and ρ
The remaining parameters are fixed at $b = 0$ and $k_1 = k_2 = 5$. Each point represents the mean over all successful runs of the experiment, i.e. over $R = 300$ when it never fails to converge. The shaded areas show the interquartile range.

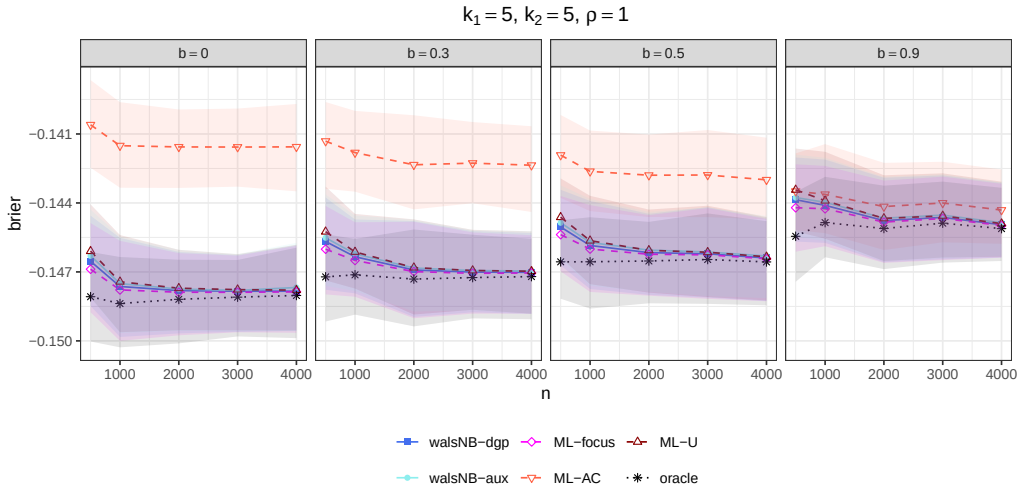

Figure S3.4: Mean validation Brier score and quartiles varying n and b
The remaining parameters are fixed at $\rho = 1$ and $k_1 = k_2 = 5$. Each point represents the mean over all successful runs of the experiment, i.e. over $R = 300$ when it never fails to converge. The shaded areas show the interquartile range.

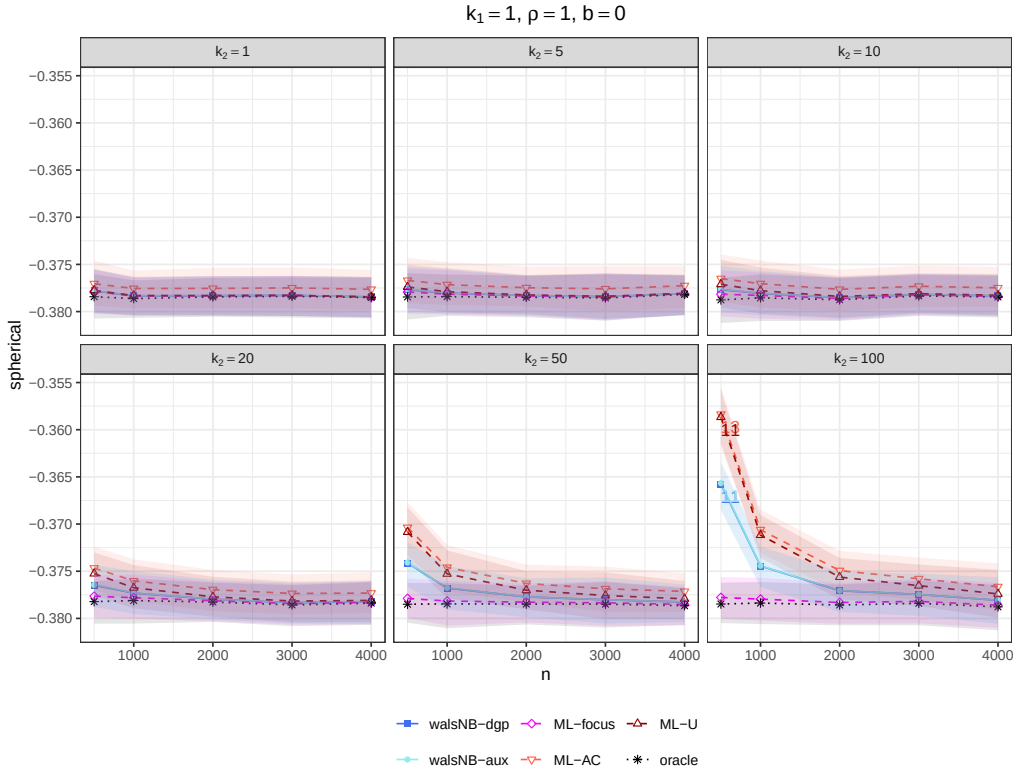

Figure S3.5: Mean validation spherical score and quartiles varying n and k_2
The remaining parameters are fixed at $k_1 = 1, \rho = 1$ and $b = 0$. Each point represents the mean over all successful runs of the experiment, i.e. over $R = 300$ when it never fails to converge. The shaded areas show the interquartile range. The number below a point indicates how often the method failed to converge in this particular setting.

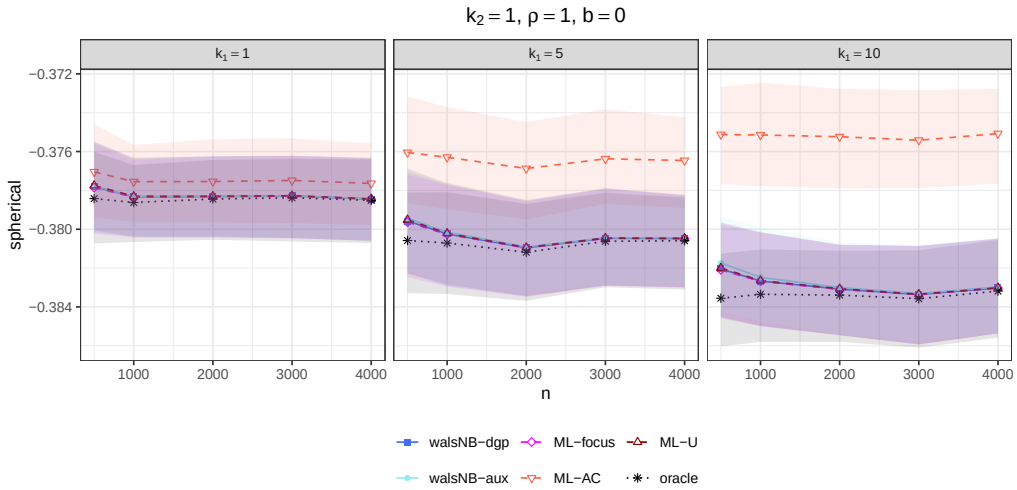

Figure S3.6: Mean validation spherical score and quartiles varying n and k_1
The remaining parameters are fixed at $k_2 = 1, \rho = 1$ and $b = 0$. Each point represents the mean over all successful runs of the experiment, i.e. over $R = 300$ when it never fails to converge. The shaded areas show the interquartile range.

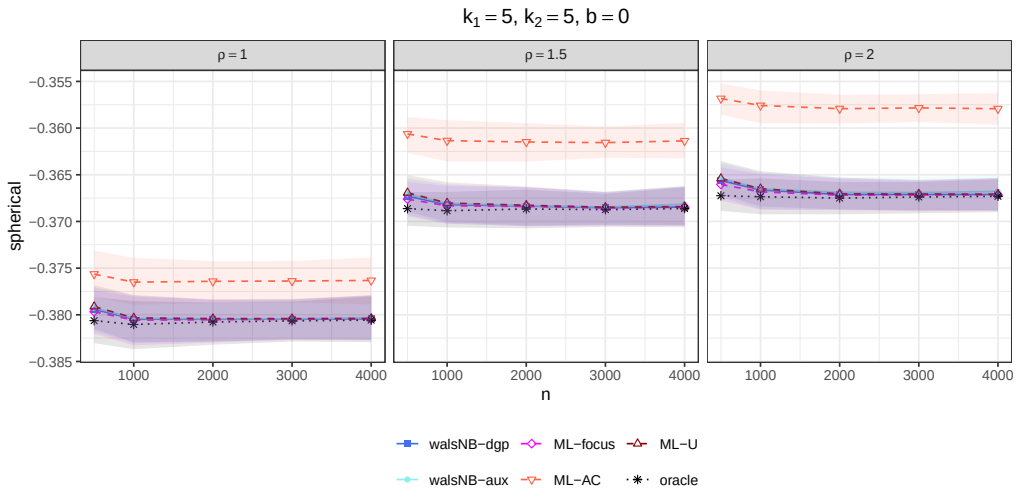

Figure S3.7: Mean validation spherical score and quartiles varying n and ρ
The remaining parameters are fixed at $b = 0$ and $k_1 = k_2 = 5$. Each point represents the mean over all successful runs of the experiment, i.e. over $R = 300$ when it never fails to converge. The shaded areas show the interquartile range.

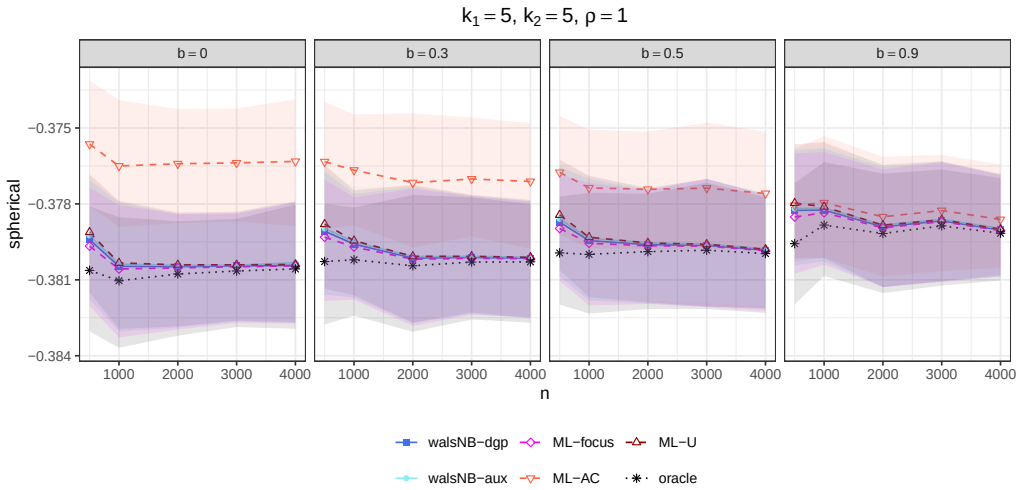

Figure S3.8: Mean validation spherical score and quartiles varying n and b
The remaining parameters are fixed at $\rho = 1$ and $k_1 = k_2 = 5$. Each point represents the mean over all successful runs of the experiment, i.e. over $R = 300$ when it never fails to converge. The shaded areas show the interquartile range.

S4 Additional results for the empirical illustration

S4.1 Results for alternative scoring rules

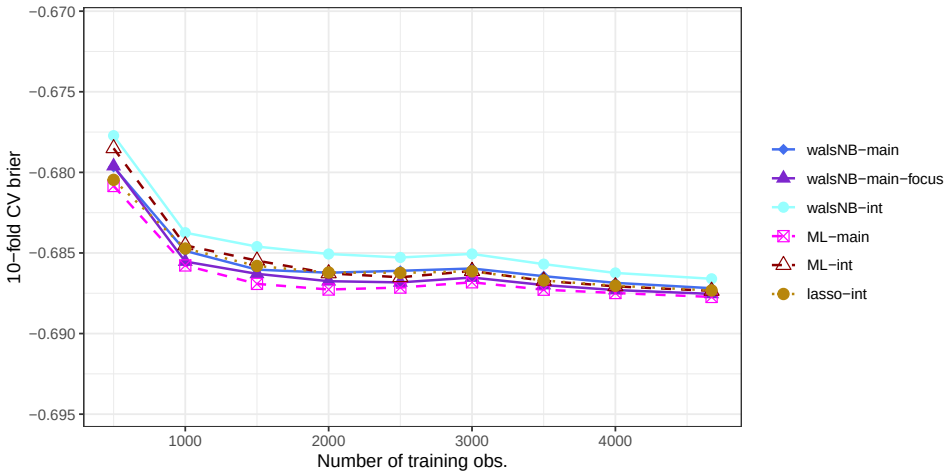

Figure S4.1: 10-fold CV Brier score varying t_l, DoctorVisits

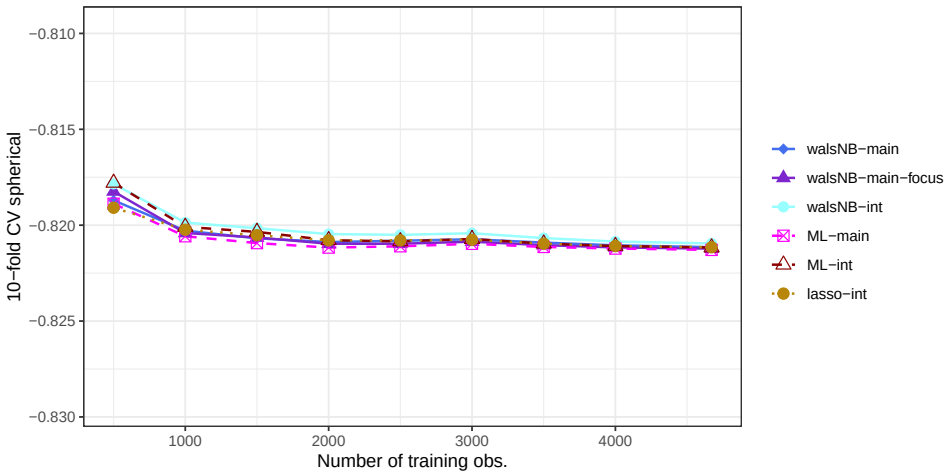

Figure S4.2: 10-fold CV spherical score varying t_l, DoctorVisits

Table S4.1: 10-fold CV Brier score varying t_l, DoctorVisits

Training obs. t_l	500	1000	1500	2000	2500	3000	3500	4000	4671
walsNB-main	-0.680	-0.685	-0.686	-0.686	-0.686	-0.686	-0.686	-0.687	-0.687
walsNB-main-focus	-0.680	-0.686	-0.686	-0.687	-0.687	-0.687	-0.687	-0.687	-0.688
walsNB-int	-0.678	-0.684	-0.685	-0.685	-0.685	-0.685	-0.686	-0.686	-0.687
ML-main	-0.681	-0.686	-0.687	-0.687	-0.687	-0.687	-0.687	-0.687	-0.688
ML-int	-0.679	-0.685	-0.685	-0.686	-0.687	-0.686	-0.687	-0.687	-0.687
lasso-int	-0.680	-0.685	-0.686	-0.686	-0.686	-0.686	-0.687	-0.687	-0.687

⁻ All figures rounded to three decimal places.

Table S4.2: 10-fold CV spherical score varying t_l, DoctorVisits

Training obs. t_l	500	1000	1500	2000	2500	3000	3500	4000	4671
walsNB-main	-0.819	-0.820	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821
walsNB-main-focus	-0.818	-0.820	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821
walsNB-int	-0.818	-0.820	-0.820	-0.820	-0.821	-0.820	-0.821	-0.821	-0.821
ML-main	-0.819	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821
ML-int	-0.818	-0.820	-0.820	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821
lasso-int	-0.819	-0.820	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821	-0.821

⁻ All figures rounded to three decimal places.

References

Abadie, A., Kasy, M., 2019. Choosing among regularized estimators in empirical economics: The risk of machine learning. Review of Economics and Statistics 101, 743–762. doi:10.1162/rest_a_00812.

Ando, T., Li, K.C., 2014. A model-averaging approach for high-dimensional regression. Journal of the American Statistical Association 109, 254–265. doi:10.1080/01621459.2013.838168.

Arnold, J.B., 2021. **ggthemes**: Extra Themes, Scales and Geoms for 'ggplot2'. URL: https://CRAN.R-project.org/package=ggthemes. R package version 4.2.4.

Boshnakov, G.N., 2023. **Rdpack**: Update and manipulate Rd documentation objects. doi:10.5281/zenodo.3925612. R package version 2.5.

Cameron, A.C., Trivedi, P.K., 1986. Econometric models based on count data. Comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1, 29–53. doi:10.1002/jae.3950010104.

Cameron, A.C., Trivedi, P.K., Milne, F., Piggott, J., 1988. A microeconometric model of the demand for health care and health insurance in Australia. The Review of Economic Studies 55, 85–106. doi:10.2307/2297531.

Crowder, M.J., 1976. Maximum likelihood estimation for dependent observations. Journal of the Royal Statistical Society: Series B (Methodological) 38, 45–53. doi:10.1111/j.2517-6161.1976.tb01565.x.

Czado, C., Gneiting, T., Held, L., 2009. Predictive model assessment for count data. Biometrics 65, 1254–1261. doi:10.1111/j.1541-0420.2009.01191.x.

Dahl, D.B., Scott, D., Roosen, C., Magnusson, A., Swinton, J., 2019. **xtable**: Export Tables to LaTeX or HTML. URL: https://CRAN.R-project.org/package=xtable. R package version 1.8-4.

De Luca, G., Magnus, J.R., 2011. Bayesian model averaging and weighted-average least squares: Equivariance, stability, and numerical issues. The Stata Journal 11, 518–544. doi:10.1177/1536867X1201100402.

De Luca, G., Magnus, J.R., Peracchi, F., 2018. Weighted-average least squares estimation of generalized linear models. Journal of Econometrics 204, 1–17. doi:10.1016/j.jeconom.2017.12.007.

De Luca, G., Magnus, J.R., Peracchi, F., 2022. Sampling properties of the Bayesian posterior mean with an application to WALS estimation. Journal of Econometrics 230, 299–317. doi:10.1016/j.jeconom.2021.04.008.

De Luca, G., Magnus, J.R., Peracchi, F., 2023. Weighted-average least squares (WALS): Confidence and prediction intervals. Computational Economics 61. doi:10.1007/s10614-022-10255-5.

Deb, P., Trivedi, P.K., 2002. The structure of demand for health care: latent class versus two-part models. Journal of Health Economics 21, 601–625. doi:10.1016/S0167-6296(02)00008-5.

Faber, F.A., Christensen, A.S., von Lilienfeld, O.A., 2020. Quantum machine learning with response operators in chemical compound space, in: Schütt, K.T., Chmiela, S., von Lilienfeld, O.A., Tkatchenko, A., Tsuda, K., Müller, K.R. (Eds.), Machine Learning Meets Quantum Physics. Springer-Verlag, Cham. chapter 8, pp. 155–169. doi:10.1007/978-3-030-40245-7_8.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression: Models, Methods and Applications. Springer-Verlag, Berlin, Heidelberg. doi:10.1007/978-3-642-34333-9.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378. doi:10.1198/016214506000001437.

Greene, W., 2008. Functional forms for the negative binomial model for count data. Economics Letters 99, 585–590. doi:10.1016/j.econlet.2007.10.015.

Heumann, C., Grenke, M., 2010. An efficient model averaging procedure for logistic regression models using a Bayesian estimator with Laplace prior, in: Kneib, T., Tutz, G. (Eds.), Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir. Physica-Verlag, Heidelberg, pp. 79–90. doi:10.1007/978-3-7908-2413-1_5.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–899. doi:10.1198/016214503000000828.

Hlavac, M., 2022. **stargazer**: Well-Formatted Regression and Summary Statistics Tables. Social Policy Institute. Bratislava, Slovakia. URL: https://CRAN.R-project.org/package=stargazer. R package version 5.2.3.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial (with discussion). Statistical Science 14, 382–417. doi:10.1214/ss/1009212519. corrected version available at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Hofner, B., Mayr, A., Robinzonov, N., Schmid, M., 2014. Model-based boosting in R: A hands-on tutorial using the R package **mboost**. Computational Statistics 29, 3–35. doi:10.1007/s00180-012-0382-5.

Hothorn, T., Leisch, F., Zeileis, A., Hornik, K., 2005. The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics 14, 675–699. doi:10.1198/106186005X59630.

Huynh, K., 2023. **WALS**: Weighted-average least squares model averaging in R. University of Basel. Mimeo.

Kleiber, C., Zeileis, A., 2008. Applied Econometrics with R. Springer-Verlag, New York. URL: https://CRAN.R-project.org/package=AER, doi:10.1007/978-0-387-77318-6.

Kolassa, S., 2016. Evaluating predictive count data distributions in retail sales forecasting. International Journal of Forecasting 32, 788–803. doi:10.1016/j.ijforecast.2015.12.004.

Lawless, J.F., 1987. Negative binomial and mixed Poisson regression. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 15, 209–225.

Madigan, D., Raftery, A.E., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association 89, 1535–1546. doi:10.1080/01621459.1994.10476894.

Magnus, J.R., De Luca, G., 2016. Weighted-average least squares (WALS): A survey. Journal of Economic Surveys 30, 117–148. doi:10.1111/joes.12094.

Magnus, J.R., Powell, O., Prüfer, P., 2010. A comparison of two model averaging techniques with an application to growth empirics. Journal of Econometrics 154, 139–153. doi:10.1016/j.jeconom.2009.07.004.

Meek, C., Thiesson, B., Heckerman, D., 2002. The learning-curve sampling method applied to model-based clustering. Journal of Machine Learning Research 2, 397–418.

Meschiari, S., 2022. **latex2exp**: Use LaTeX Expressions in Plots. URL: https://CRAN.R-project.org/package=latex2exp. R package version 0.9.6.

Min, C., Zellner, A., 1993. Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. Journal of Econometrics 56, 89–118. doi:10.1016/0304-4076(93)90102-B.

Mullahy, J., 1997. Heterogeneity, excess zeros, and the structure of count data models. Journal of Applied Econometrics 12, 337–350. doi:10.1002/(SICI)1099-1255(199705)12:3<337::AID-JAE438>3.0.CO;2-G.

Newey, W.K., McFadden, D.L., 1994. Large sample estimation and hypothesis testing, in: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics. North-Holland, Amsterdam. volume 4. chapter 36, pp. 2111–2245.

Plate, T., Heiberger, R., 2016. **abind**: Combine Multidimensional Arrays. URL: https://CRAN.R-project.org/package=abind. R package version 1.4-5.

Raftery, A.E., Hoeting, J.A., Volinsky, C.T., Painter, I., Yeung, K.Y., 2020. **BMA**: Bayesian Model Averaging. URL: https://CRAN.R-project.org/package=BMA. R package version 3.18.12.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rupp, M., Tkatchenko, A., Müller, K.R., von Lilienfeld, O.A., 2012. Fast and accurate modeling of molecular atomization energies with machine learning. Physical Review Letters 108, 058301. doi:10.1103/PhysRevLett.108.058301.

Steel, M.F.J., 2020. Model averaging and its use in economics. Journal of Economic Literature 58, 644–719. doi:10.1257/jel.20191385.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Statistics and Computing. 4th ed., Springer-Verlag, New York. URL: https://www.stats.ox.ac.uk/pub/MASS4/, doi:10.1007/978-0-387-21706-2.

Wang, Z., 2023. **mpath**: Regularized Linear Models. URL: https://CRAN.R-project.org/package=mpath. R package version 0.4-2.23.

Wang, Z., Ma, S., Zappitelli, M., Parikh, C., Wang, C.Y., Devarajan, P., 2016. Penalized count data regression with application to hospital stay after pediatric cardiac surgery. Statistical Methods in Medical Research 25, 2685–2703. doi:10.1177/0962280214530608.

Wickham, H., 2016. **ggplot2**: Elegant Graphics for Data Analysis. 2nd ed., Springer-Verlag, New York. URL: https://ggplot2.tidyverse.org, doi:10.1007/978-3-319-24277-4.

Winkler, R.L., 1996. Scoring rules and the evaluation of probabilities. Test 5, 1–60. doi:10.1007/BF02562681.

Zeileis, A., Croissant, Y., 2010. Extended model formulas in R: Multiple parts and multiple responses. Journal of Statistical Software 34, 1–13. doi:10.18637/jss.v034.i01.

Zhang, X., Liu, C.A., 2019. Inference after model averaging in linear regression models. Econometric Theory 35, 816–841. doi:10.1017/S0266466618000269.

Zhang, X., Yu, D., Zou, G., Liang, H., 2016. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. Journal of the American Statistical Association 111, 1775–1790. doi:10.1080/01621459.2015.1115762.

Zhang, X.D., 2017. Matrix Analysis and Applications. Cambridge University Press, Cambridge. doi:10.1017/9781108277587.