Separating the Wheat from the Chaff: Fast Estimation of GLMs with High-Dimensional Fixed Effects

Julian Hinz,[*] Alexander Hudlet,[†] and Joschka Wanner,[‡]

March 21, 2019

Abstract

In this note we describe how the first-order conditions of fixed effects in certain generalized linear models display a very convenient form. This can be exploited to implement a very fast GLM estimator with high-dimensional fixed effects, as we do in the `R` package `glmhdfe`.

WORK IN PROGRESS

Keywords: generalized linear model, high-dimensional fixed effects, gravity
JEL Classification: F12, F14, F15

[*]European University Institute and Kiel Centre for Globalization
[†]University of Bayreuth and Kiel Institute for the World Economy
[‡]University of Bayreuth

1 Introduction

We introduce a new GLM estimator implementation with high dimensional fixed effects. The note is structured as follows. In section 2 we derive the first order conditions for the generalized linear model yielding a convenient closed form for the fixed effects — given other parameter estimates — for certain family-link function combinations, followed by the description of the computation of standard errors in section 3. In section 4 we document the `R` implementation in the `glmhdfe` package.

2 GLMs with high-dimensional fixed effects

Consider the following equation that we want to estimate:

$$y_i = g^{-1}\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \left(\boldsymbol{d}_i^A\right)'\boldsymbol{\delta}^A + \left(\boldsymbol{d}_i^B\right)'\boldsymbol{\delta}^B + \ldots\right) + \epsilon_i,$$

or in matrix form

$$\begin{aligned}\boldsymbol{y} &= g^{-1}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}^A\boldsymbol{\delta}^A + \boldsymbol{D}^B\boldsymbol{\delta}^B + \ldots\right) + \boldsymbol{\epsilon} \\ &= g^{-1}(\boldsymbol{\eta}) + \boldsymbol{\epsilon},\end{aligned}$$

such that

$$\begin{aligned}E(\boldsymbol{y}) &= \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) \quad \text{and} \\ E(y_i) &= \mu_i = g^{-1}(\eta_i),\end{aligned}$$

where subscript i denotes the observation, $g(\cdot)$ is the link function, y_i denotes the dependent variable, $\boldsymbol{x}_i$ denotes the vector of explanatory variables with corresponding parameter vector $\boldsymbol{\beta}$, and $\boldsymbol{d}_i$ denoting dummies where superscripts $A, B, \ldots$ index the arbitrary number of fixed effects. $\boldsymbol{\delta}$ are the conforming parameters. ϵ_i is a remainder error term.

Using results from Nelder and Wedderburn (1972), McCullagh and Nelder (1989), and specifically generalizing the way to write the first order conditions by Egger and Staub (2016) to a general context with arbitrary fixed effects, we get the following first-order conditions:

$$\widehat{\boldsymbol{\beta}}: \quad \sum_i \frac{y_i - \mu_i}{V(y_i)} \frac{\partial \mu_i}{\partial \widehat{\boldsymbol{\beta}}} = \boldsymbol{0}, \tag{1a}$$

$$\widehat{\delta}^a: \quad \sum_i \frac{y_i - \mu_i}{V(y_i)} \frac{\partial \mu_i}{\partial \widehat{\delta}^a} = 0, \tag{1b}$$

$$\ldots$$

where $\widehat{\delta}^a$ is the ath element of $\widehat{\boldsymbol{\delta}}^A$.

Importantly, note that the inner derivative of μ_i with respect to $\widehat{\delta}^a$ is always equal to d_i^a, which is the ath element of $\boldsymbol{d}_i^A$. Therefore, in the corresponding first-order conditions, it will always suffice to consider the elements of the sum for which $d_i^a = 1$, i.e. equation (1b) can equivalently be written as:

$$\widehat{\delta}^a: \quad \sum_{i|d_i^a=1} \frac{y_i - \mu_i}{V(y_i)} = 0. \qquad \text{(1b)}'$$

This property makes the computation of the fixed effects vector $\widehat{\delta}^a$ dramatically less costly, as the estimation is simply a collection of summation operations.

Following Nelder and Wedderburn (1972), $\widehat{\boldsymbol{\beta}}$ can be obtained using IRLS (while updating (1b)′ in each iteration) by:

$$\begin{aligned} &\widehat{\boldsymbol{\beta}} = \left[\boldsymbol{X'WX}\right]^{-1} \boldsymbol{X'W}\widetilde{\boldsymbol{y}}, \\ \text{with} \quad &\widetilde{y}_i = (y_i - \widehat{\mu}_i)\frac{\partial \widehat{\eta}_i}{\partial \widehat{\mu}_i} + \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \quad \text{and} \quad w_{ii} = V(\widehat{\mu}_i)^{-1}\left(\frac{\partial \widehat{\mu}_i}{\partial \widehat{\eta}_i}\right)^2. \end{aligned} \qquad (2)$$

Note that $\mathbf{W}$ is a diagonal weighting matrix and hence w_{ii} refers to the elements of the main diagonal of $\mathbf{W}$. $\widehat{\mu}_i$ and $\widehat{\eta}_i$ denote the *current* value in the iterative procedure.

We now explore equations (1a), (1b)' and (2) for some specific family and link function.

2.1 Gaussian with identity link (OLS)

In the case of the Gaussian family the variance is given by $\mathrm{V}(\mu) = 1$. Together with an identity link this yields the following FOCs:

$$\widehat{\boldsymbol{\beta}}: \quad \sum_i \left(y_i - \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)\right)\boldsymbol{x}_i = \mathbf{0}, \qquad (3a)$$

$$\widehat{\delta}^a: \quad \sum_i \left(y_i - \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)\right)d_i^a = 0, \qquad (3b)$$

or, equivalently,

$$\widehat{\delta}^a: \quad \sum_{i|d_i^a=1} \left(y_i - \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)\right) = 0, \qquad (3c)$$

$$\ldots$$

An iteration then consists of evaluating the following equations:

$$\widehat{\boldsymbol{\beta}} = \left[\boldsymbol{X'WX}\right]^{-1}\boldsymbol{X'W}\widetilde{\boldsymbol{y}}, \quad \text{with} \quad \widetilde{y}_i = y_i - \widehat{\mu}_i + \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \quad \text{and} \quad w_{ii} = 1, \tag{3d}$$

$$\widehat{\delta}^a = \frac{1}{n^a}\sum_{i|d_i^a=1}\left(y_i - \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)\right), \tag{3e}$$

$$\ldots$$

where $n^a = \sum_i d_i^a$.

2.2 Gaussian with log link (NLS)

Combining a Gaussian family with a log link and $\mathrm{V}(\mu) = 1$ yields the following FOCs:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}: \quad & \sum_i \left(y_i - \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)\right) \\
& \qquad \times \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)\boldsymbol{x}_i = \boldsymbol{0}, \qquad (4a) \\
\widehat{\delta}^a: \quad & \sum_{i|d_i^a=1} \left(y_i - \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)\right) \\
& \qquad \times \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right) = 0, \qquad (4b) \\
& \ldots
\end{aligned}$$

An iteration then consists of evaluating the following equations:

$$\widehat{\boldsymbol{\beta}} = \left[\boldsymbol{X'WX}\right]^{-1}\boldsymbol{X'W}\widetilde{\boldsymbol{y}}, \quad \text{with} \quad \widetilde{y}_i = \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} + \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \quad \text{and} \quad w_{ii} = \widehat{\mu}_i^2, \tag{4c}$$

$$\widehat{\delta}^a = \log\left(\frac{\sum_{i|d_i^a=1} y_i \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)}{\sum_{i|d_i^a=1}\exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)^2}\right), \tag{4d}$$

$$\ldots$$

2.3 Poisson with log link

In the case of the Poisson family the variance is given by $\mathrm{V}(\mu) = \mu$. Together with a log link this yields the following FOCs:

$$\widehat{\boldsymbol{\beta}}: \quad \sum_i \left(y_i - \exp\left(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)' \widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)' \widehat{\boldsymbol{\delta}}^B + \ldots \right) \right) \boldsymbol{x}_i = \mathbf{0}, \tag{5a}$$

$$\widehat{\delta}^a: \quad \sum_{i|d_i^a=1} y_i - \exp\left(\boldsymbol{z}_i' \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)' \widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)' \widehat{\boldsymbol{\delta}}^B + \ldots \right) = 0, \tag{5b}$$

$$\ldots$$

An iteration then consists of evaluating the following equations:

$$\widehat{\boldsymbol{\beta}} = \left[\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right]^{-1} \boldsymbol{X}'\boldsymbol{W}\widetilde{\boldsymbol{y}}, \quad \text{with} \quad \widetilde{y}_i = \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} + \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \quad \text{and} \quad w_{ii} = \widehat{\mu}_i, \tag{5c}$$

$$\widehat{\delta}^a = \log\left(\frac{\sum_{i|d_i^a=1} y_i}{\sum_{i|d_i^a=1} \exp\left(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^B\right)' \widehat{\boldsymbol{\delta}}^B + \ldots \right)} \right), \tag{5d}$$

$$\ldots$$

2.4 Gamma with log link

In the case of the Gamma family the variance is given by $\mathrm{V}(\mu) = \mu^2$. Together with a log link this yields the following FOCs:

$$\widehat{\boldsymbol{\beta}}: \quad \sum_i \frac{y_i - \exp\left(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)' \widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)' \widehat{\boldsymbol{\delta}}^B + \ldots \right)}{\exp\left(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)' \widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)' \widehat{\boldsymbol{\delta}}^B + \ldots \right)} \boldsymbol{x}_i = \mathbf{0}, \tag{6a}$$

$$\widehat{\delta}^a: \quad \sum_{i|d_i^a=1} \frac{y_i - \exp\left(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)' \widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)' \widehat{\boldsymbol{\delta}}^B + \ldots \right)}{\exp\left(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)' \widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)' \widehat{\boldsymbol{\delta}}^B + \ldots \right)} = 0, \tag{6b}$$

$$\ldots$$

An iteration then consists of evaluating the following equations:

$$\widehat{\boldsymbol{\beta}} = \left[\boldsymbol{X}'\boldsymbol{X}\right]^{-1} \boldsymbol{X}'\widetilde{\boldsymbol{y}}, \quad \text{with} \quad \widetilde{y}_i = \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} + \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}, \tag{6c}$$

$$\widehat{\delta}^a = \log\left(\frac{1}{n^a} \sum_{i|d_i^a=1} \frac{y_i}{\exp\left(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^B\right)' \widehat{\boldsymbol{\delta}}^B + \ldots \right)} \right), \tag{6d}$$

$$\ldots$$

2.5 Inverse Gaussian

In the case of the inverse Gaussian family the variance is given by $\mathrm{V}(\mu) = \mu^3$. Together with a log link this yields the following FOCs:

$$\widehat{\boldsymbol{\beta}}: \quad \sum_i \frac{y_i - \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)}{\exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)^2}\boldsymbol{x}_i = \mathbf{0}, \tag{7a}$$

$$\widehat{\delta}^a: \quad \sum_{i|d_i^a=1} \frac{y_i - \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)}{\exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)^2} = 0, \tag{7b}$$

$$\ldots$$

An iteration then consists of evaluating the following equations:

$$\widehat{\boldsymbol{\beta}} = \left[\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{W}\widetilde{\boldsymbol{y}}, \quad \text{with} \quad \widetilde{y}_i = \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} + \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \quad \text{and} \quad w_{ii} = \widehat{\mu}_i^{-1}, \tag{7c}$$

$$\widehat{\delta}^a = \log\left(\frac{\sum_{i|d_i^a=1} y_i / \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)^2}{\sum_{i|d_i^a=1} \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)^{-1}}\right), \tag{7d}$$

$$\ldots$$

3 Standard Errors

Standard errors can be computed after convergence of the iteration described above by making use of the Frisch-Waugh-Lovell theorem in combination with a weighted pseudo-demeaning to get rid of the influence of the fixed effects, similar to Gaure (2013) and Stammann (2018). We demonstrate the procedure in the following for two sets of fixed effects, but the procedure generalizes to an arbitrary number.

The model described above can be expressed as

$$\sqrt{\boldsymbol{W}}\tilde{\boldsymbol{y}} = \sqrt{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{\beta} + \sqrt{\boldsymbol{W}}\boldsymbol{D}^A\boldsymbol{\delta}^a + \sqrt{\boldsymbol{W}}\boldsymbol{D}^B\boldsymbol{\delta}^b + \sqrt{\boldsymbol{W}}\boldsymbol{\varepsilon},$$

where $\sqrt{\boldsymbol{W}}$ is the diagonal weights matrix, $\boldsymbol{D}^A$ and $\boldsymbol{D}^B$ are the two sets of fixed effects with the corresponding parameter vectors $\boldsymbol{\delta}^a$ and $\boldsymbol{\delta}^b$.

The optimization problem is then

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta}^a,\boldsymbol{\delta}^b} S = \left(\sqrt{\boldsymbol{W}}\tilde{\boldsymbol{y}} - \sqrt{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{\beta} - \sqrt{\boldsymbol{W}}\boldsymbol{D}^A\boldsymbol{\delta}^a - \sqrt{\boldsymbol{W}}\boldsymbol{D}^B\boldsymbol{\delta}^b\right)' \left(\sqrt{\boldsymbol{W}}\tilde{\boldsymbol{y}} - \sqrt{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{\beta} - \sqrt{\boldsymbol{W}}\boldsymbol{D}^A\boldsymbol{\delta}^a - \sqrt{\boldsymbol{W}}\boldsymbol{D}^B\boldsymbol{\delta}^b\right),$$

such that the corresponding normal equations are given by

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}'\boldsymbol{W}\tilde{\boldsymbol{y}} + \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{W}\boldsymbol{D}^A\boldsymbol{\delta}^a + \boldsymbol{X}'\boldsymbol{W}\boldsymbol{D}^B\boldsymbol{\delta}^b = 0, \tag{8}$$

$$\frac{\partial S}{\partial \boldsymbol{\delta}^a} = -\boldsymbol{D}^{A\prime}\boldsymbol{W}\tilde{\boldsymbol{y}} + \boldsymbol{D}^{A\prime}\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}^{A\prime}\boldsymbol{W}\boldsymbol{D}^A\boldsymbol{\delta}^a + \boldsymbol{D}^{A\prime}\boldsymbol{W}\boldsymbol{D}^B\boldsymbol{\delta}^b = 0, \tag{9}$$

$$\frac{\partial S}{\partial \boldsymbol{\delta}^b} = -\boldsymbol{D}^{B\prime}\boldsymbol{W}\tilde{\boldsymbol{y}} + \boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{D}^A\boldsymbol{\delta}^a + \boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{D}^B\boldsymbol{\delta}^b = 0. \tag{10}$$

Solving equation (10) for $\boldsymbol{\delta}^b$ yields

$$\boldsymbol{\delta}^b = (\boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{D}^B)^{-1}(\boldsymbol{D}^{B\prime}\boldsymbol{W}\tilde{\boldsymbol{y}} - \boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{D}^A\boldsymbol{\delta}^a), \tag{11}$$

which, putting it back into equation (8), yields

$$\begin{aligned}
&\boldsymbol{X}'\boldsymbol{W}\underbrace{(\boldsymbol{I} - \boldsymbol{D}^B(\boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{D}^B)^{-1}\boldsymbol{D}^{B\prime}\boldsymbol{W})}_{:=\boldsymbol{P}^B}\boldsymbol{X}\boldsymbol{\beta} + \\
&\boldsymbol{X}'\boldsymbol{W}\underbrace{(\boldsymbol{I} - \boldsymbol{D}^B(\boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{D}^B)^{-1}\boldsymbol{D}^{B\prime}\boldsymbol{W})}_{:=\boldsymbol{P}^B}\boldsymbol{D}^A\delta^a \\
&= \boldsymbol{X}'\boldsymbol{W}\underbrace{(\boldsymbol{I} - \boldsymbol{D}^B(\boldsymbol{D}^{B\prime}\boldsymbol{W}\boldsymbol{D}^B)^{-1}\boldsymbol{D}^{B\prime}\boldsymbol{W})}_{:=\boldsymbol{P}^B}\tilde{\boldsymbol{y}} \\
&\Leftrightarrow \boldsymbol{X}'\boldsymbol{W}\boldsymbol{P}^B\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{W}\boldsymbol{P}^B\boldsymbol{D}^A\boldsymbol{\delta}^a = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{P}^B\tilde{\boldsymbol{y}} \\
&\Leftrightarrow (\boldsymbol{P}^B\boldsymbol{X})'\boldsymbol{W}(\boldsymbol{P}^B\boldsymbol{X})\boldsymbol{\beta} + (\boldsymbol{P}^B\boldsymbol{X})'\boldsymbol{W}(\boldsymbol{P}^B\boldsymbol{D}^A)\boldsymbol{\delta}^a = (\boldsymbol{P}^B\boldsymbol{X})'\boldsymbol{W}(\boldsymbol{P}^B\tilde{\boldsymbol{y}}) \\
&\Leftrightarrow \dot{\boldsymbol{X}}'\boldsymbol{W}\dot{\boldsymbol{X}}\boldsymbol{\beta} + \dot{\boldsymbol{X}}'\boldsymbol{W}\dot{\boldsymbol{D}}^A\boldsymbol{\delta} = \dot{\boldsymbol{X}}'\boldsymbol{W}\dot{\tilde{\boldsymbol{y}}},
\end{aligned} \tag{12}$$

where $\boldsymbol{P}^B$ is the projection matrix for $\boldsymbol{D}^B$ and the "$\cdot$" superscript denotes a pseudo-demeaning over one dimension. Analogously inserting equation (11) into equation (9) yields

$$\begin{aligned}
\Leftrightarrow (\boldsymbol{P}^B\boldsymbol{D}^A)'\boldsymbol{W}(\boldsymbol{P}^B\boldsymbol{X})\boldsymbol{\beta} + (\boldsymbol{P}^B\boldsymbol{D}^A)'\boldsymbol{W}(\boldsymbol{P}^B\boldsymbol{D}^A)\boldsymbol{\delta}^a &= (\boldsymbol{P}^B\boldsymbol{D}^A)'\boldsymbol{W}(\boldsymbol{P}^B\tilde{\boldsymbol{y}}) \\
\Leftrightarrow \dot{\boldsymbol{D}}^{A\prime}\boldsymbol{W}\dot{\boldsymbol{D}}^A\boldsymbol{\delta} + \dot{\boldsymbol{D}}^{A\prime}\boldsymbol{W}\dot{\boldsymbol{X}}\boldsymbol{\beta} &= \dot{\boldsymbol{D}}^{A\prime}\boldsymbol{W}\dot{\tilde{\boldsymbol{y}}}.
\end{aligned} \tag{13}$$

Solving equation (13) for $\boldsymbol{\delta}^a$ yields

$$\Leftrightarrow \boldsymbol{\delta}^a = (\dot{\boldsymbol{D}}^A\boldsymbol{W}\dot{\boldsymbol{D}}^A)^{-1}(\dot{\boldsymbol{D}}^{A\prime}\boldsymbol{W}\dot{\tilde{\boldsymbol{y}}} - \dot{\boldsymbol{D}}^{A\prime}\boldsymbol{W}\dot{\boldsymbol{X}}\boldsymbol{\beta}). \tag{14}$$

Finally inserting equation (14) back into (13) yields

$$\begin{aligned}
\Leftrightarrow \dot{\boldsymbol{X}}'\boldsymbol{W}\underbrace{(\boldsymbol{I}-\dot{\boldsymbol{D}}^A(\dot{\boldsymbol{D}}^{A\prime}\boldsymbol{W}\dot{\boldsymbol{D}}^A)^{-1}\dot{\boldsymbol{D}}^A\boldsymbol{W})}_{:=\boldsymbol{P}^A}\dot{\boldsymbol{X}}\boldsymbol{\beta} &= \dot{\boldsymbol{X}}'\boldsymbol{W}\underbrace{(\boldsymbol{I}-\dot{\boldsymbol{D}}^A(\dot{\boldsymbol{D}}^{A\prime}\boldsymbol{W}\dot{\boldsymbol{D}}^A)^{-1}\dot{\boldsymbol{D}}^A\boldsymbol{W})}_{:=\boldsymbol{P}^A}\dot{\tilde{\boldsymbol{y}}} \\
\Leftrightarrow (\boldsymbol{P}^A\dot{\boldsymbol{X}})'\boldsymbol{W}(\boldsymbol{P}^A\dot{\boldsymbol{X}})\boldsymbol{\beta} &= (\boldsymbol{P}^A\dot{\boldsymbol{X}})'\boldsymbol{W}(\boldsymbol{P}^A\dot{\tilde{\boldsymbol{y}}}) \\
\Leftrightarrow (\boldsymbol{P}^A(\boldsymbol{P}^B\boldsymbol{X}))'\boldsymbol{W}(\boldsymbol{P}^A(\boldsymbol{P}^B\boldsymbol{X}))\boldsymbol{\beta} &= (\boldsymbol{P}^A(\boldsymbol{P}^B\boldsymbol{X}))'\boldsymbol{W}(\boldsymbol{P}^A(\boldsymbol{P}^B\tilde{\boldsymbol{y}})) \\
\Leftrightarrow \ddot{\boldsymbol{X}}'\boldsymbol{W}\ddot{\boldsymbol{X}}\boldsymbol{\beta} &= \ddot{\boldsymbol{X}}'\boldsymbol{W}\ddot{\tilde{\boldsymbol{y}}}, \qquad (15)
\end{aligned}$$

where $\boldsymbol{P}^A$ is the projection matrix for $\boldsymbol{D}^A$ and the "$\cdot\cdot$" superscript denotes a pseudo-demeaning over both dimension.

In practice, the demeaning can be achieved via a simple iteration procedure. Call $\overline{m}$ the pseudo-demeaned transformation of m and n the iteration. Then, starting with an initial $\overline{m}_i^0 = m_i$, iterate over steps (1), (2), ... until $\overline{m}_i^{(n)} - \overline{m}_i^{(n-1)} \leq \varkappa$

$$\begin{aligned}
&(1) \quad \overline{m}_i^{(n)} = \overline{m}_i^{(n-1)} - \frac{\sum_{i|d_i^a=1} w_i \cdot \overline{m}_i^{(n-1)}}{\sum_{i|d_i^a=1} w_i} \\
&(2) \quad \overline{m}_i^{(n+1)} = \overline{m}_i^{(n)} - \frac{\sum_{i|d_i^b=1} w_i \cdot \overline{m}_i^{(n)}}{\sum_{i|d_i^b=1} w_i} \\
&\ldots
\end{aligned}$$

so that $\overline{m}_i \approx \ddot{m}_i$.

Using the demeaned $\ddot{\boldsymbol{X}}$ matrix, standard errors can be computed from the inverse Hessian

$$\ddot{\boldsymbol{H}}^{-1} = \gamma\hat{\phi}\left(\ddot{\boldsymbol{X}}'\ddot{\boldsymbol{X}}\right)^{-1}$$

where γ is the appropriate degree-of-freedom adjustment and $\hat{\phi}$ is either the MM estimate of the scale parameter or one.[1]

This inverse Hessian is also used to compute the robust "sandwich" errors, using the pseudo-demeaned score vector $\ddot{\boldsymbol{\xi}}$, where

$$\xi_i = (y_i - \widehat{\mu}_i)\frac{\partial\widehat{\eta}_i}{\partial\widehat{\mu}_i}.$$

Robust standard errors are obtained by column-wise multiplication of $\ddot{\boldsymbol{X}}$ with $\ddot{\boldsymbol{\xi}}$ to obtain

[1] $\phi = 1$ for Poisson, Bernoulli, binomial and negative-binomial distributions. ϕ has to be estimated for Gaussian, gamma, and inverse-Gaussian distributions, where $\hat{\phi}$ is given by: $\widehat{\phi} = \frac{1}{N-K}\sum_i = 1^N \frac{(y_i-\mu_i)^2}{V(\mu_i)}$

the gradient $\ddot{g}$, and then multiplying

$$\ddot{H}^{-1}\ddot{g}'\ddot{g}\ddot{H}^{-1}.$$

Clustered standard errors are easily obtained by summing the gradient $\ddot{g}$ appropriately beforehand.

4 Implementation in R and Example

We implement the procedure in R in the package glmhdfe. The package can be installed from Github via the remotes package:

```
remotes::install_github("julianhinz/R_glmhdfe")
```

or by downloading and installing the zipped releases from https://github.com/julianhinz/R_glmhdfe/releases. The glmhdfe function has a similar syntax as the felm function from the lfe package and the feglm function in the alpaca package:

```
glmhdfe(trade ~ fta | iso_o_year + iso_d_year + iso_o_iso_d | iso_o
↪  + iso_d + year,
    family = poisson(link = "log"),
    data = data)
```

The first part of the formula is specified as usual. The second part of the formula specifies the fixed effects dimensions, the third part, which is optional, the clustering of the standard errors.

4.1 Options

There are numerous options to tweak the estimation procedure:

- `formula` describes dependent variable, right-hand side variables of interest, sets of fixed effects and clustering of standard errors, e.g. as `y ~ x | fe1 + fe2 | cluster1 + cluster2`
- `data` specifies the data.table or data.frame with data used in the regression
- `family` specifies the estimator used, currently limited to `gaussian(link = "identity")`, `gaussian(link = "log")`, `poisson(link = "log")`, `Gamma(link = "log")`
- `beta` allows to include a vector of starting values, although, interestingly, this does not tend to speed up the estimation

- `tolerance` specifies the minimum change in the deviance at which the iteration breaks
- `max_iterations` specifies the maximum number of iterations
- `accelerate` specifies whether to use an acceleration algorithm, still quite buggy
- `accelerate_iterations` specifies the number of iterations before starting acceleration algorithm
- `accelerate_aux_vector` specifies whether to include the estimated fixed effects vectors in IRLS, which, interestingly, increases convergence speed
- `compute_vcov` asks whether to compute the variance-covariance matrix. It can also be computed ex-post when data from estimation provided
- `demean_variables` if you don't want to compute the variance-covariance matrix right away, do you still want to demean variables to be used in estimation of variance-covariance matrix?
- `demean_iterations` specifies the number of iterations for the demeaning
- `demean_tolerance` specifies the minimum change in the diagonal of the Hessian at which the demeaning iteration breaks
- `include_fe` asks whether the estimated fixed effects should be returned
- `include_data` asks whether the data used in the estimation should be returned, which may be useful if the variance-covariance matrix will be computed ex-post
- `include_data_vcov` return data used in variance-covariance matrix estimation?
- `skip_checks` specifies, whether certain data integrity checks should be skipped before starting the procedure. Current option to skip are the detection of separation issues (`"separation"`), multicollinearity (`"multicollinearity"`), or missing data (`"complete_cases"`)
- `trace` asks whether to show some information during the estimation
- `verbose` asks whether to show a bit more information during estimation for the impatient

4.2 Other functions

There is also the usual battery of generic functions, like `coef` , `summary` , etc. Furthermore, if for some reason you want to (re-)estimate the variance-covariance matrix afterwards, or change the level of clustering, you can do so with the `compute_vcov` command:

```
compute_vcov(data, call, info)
```

You need to specify the `data` (best in the form of a `glmhdfe_data` object), `call` (for information on clustering and variable of interest), and `info` (for information on degrees of freedom, etc.).

4.3 Speed

One advantage of separating the updating of the fixed effects from that of the other right-hand side variables is computation time. Figures 1, 2 and 3 show the time until convergence in seconds plotted against the number of observations in a setting with two or three sets of fixed effects. We do so in a scenario of fitting a standard model from the literature in international trade with a Poisson estimator. We generate the data using a so-called structural gravity equation of international trade that relates bilateral flows to exporter-specific, importer-specific and bilateral determinants.[2] The number of observations is determined by the fixed effects dimensions of the data. In figure 1 we assume no time dimension (and hence two sets of fixed effects) and increase the number of origin and destinations countries at the same time from 4 to 400, generating samples with between $4^2 = 16$ and $400^2 = 160,000$ observations and up to $2 \times 400 = 800$ fixed effects. While the `base R` command `glm` takes more than 10 minutes to converge for the largest of the samples, `glmhdfe` still takes only around 0.8 seconds for the same task. In figures 2 and 3 we increase the sample size and number of fixed effects dramatically to showcase the speed advantages of `glmhdfe`. We introduce a time dimension t fixed at 100 and increase the number of o and d at the same time from 60 to 600. This generates samples between $360,000$ observations with $2 \times 60 \times 100 + 60 \times 60 = 15,600$ fixed effects ($12,000$ in the two way case), and 36,000,000 observations with $2 \times 600 \times 100 + 600 \times 600 = 480,000$ fixed effects ($120,000$ in the two way case). The figures compare the performance of different options of the `glmhdfe` command.

In figures 4 and 5 we plot the seconds until convergence against the number of observations and the number of fixed effect groups, the respective other held constant. Figure 4 holds the number of observations fixed at 1,000,000, with the number of fixed effect groups increasing from 2 to 500,000. Figure 5 holds the number of fixed effects constant at 50,000, but increases the number of observations from 100,000 to 5,000,000 (with *roughly* equal number of observations within groups).

[2] See Head and Mayer (2014) for details.

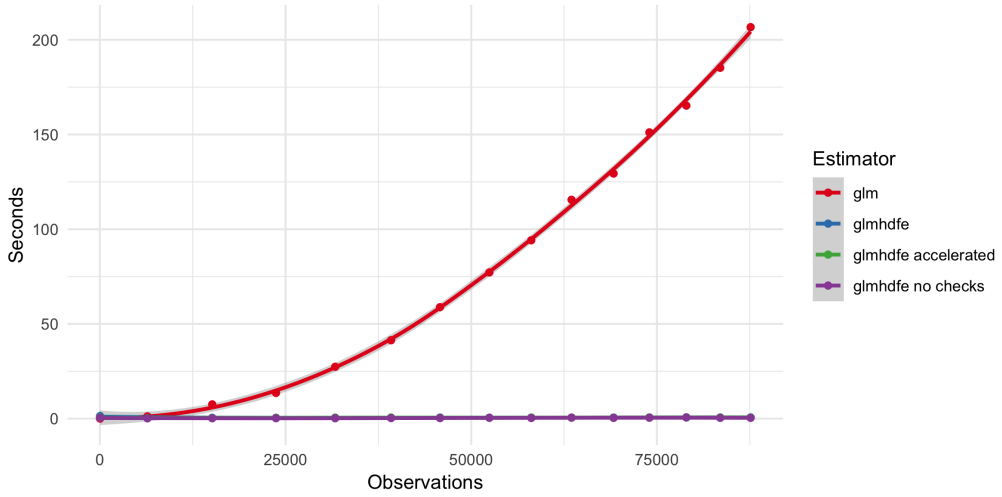

Figure 1: Time until convergence with two-way fixed effects

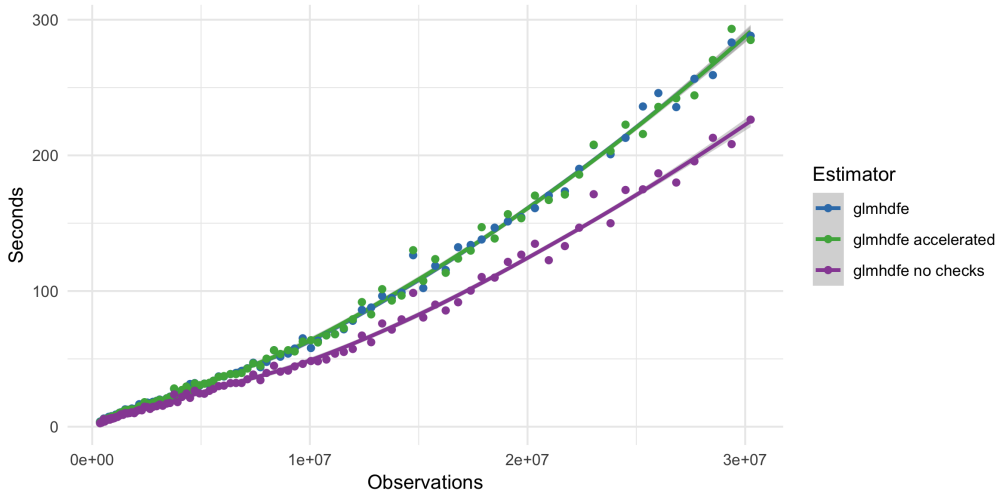

Figure 2: Time until convergence with two-way fixed effects

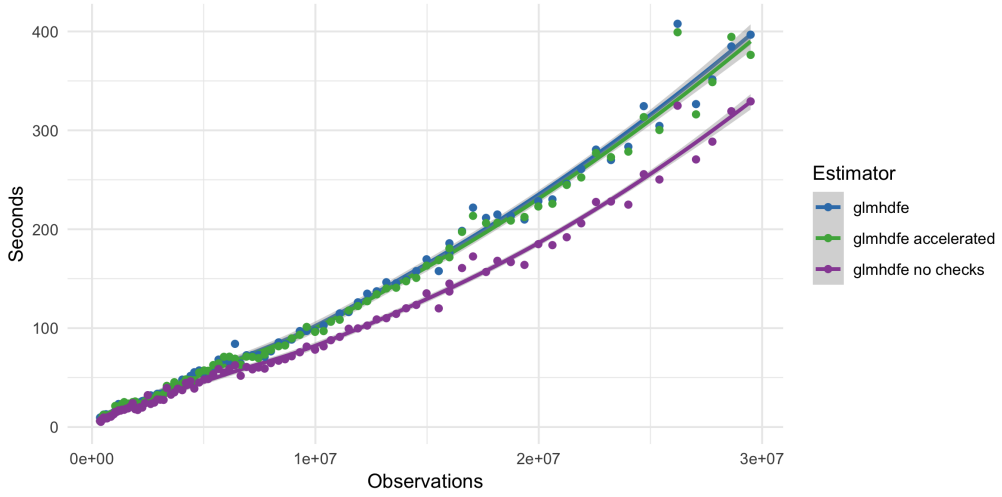

Figure 3: Time until convergence with three-way fixed effects

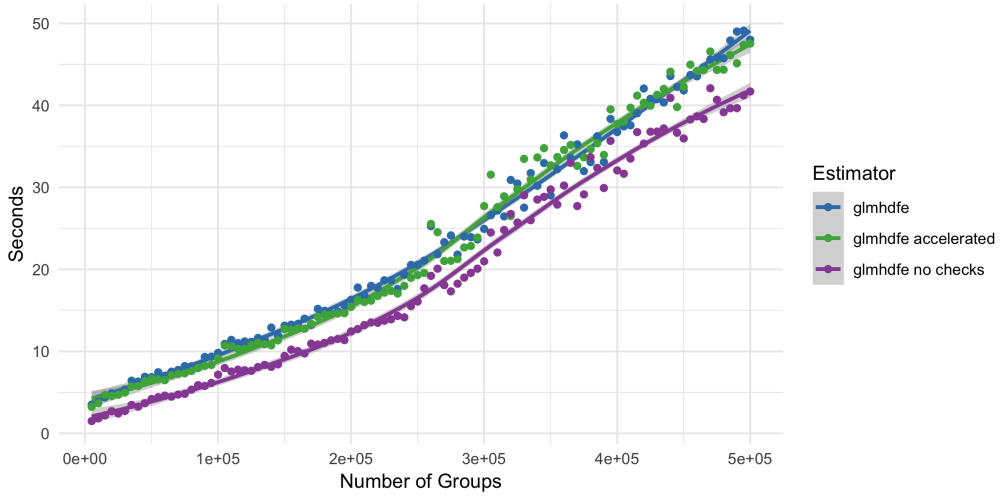

Figure 4: Time until convergence with fixed number of observations and increasing number groups

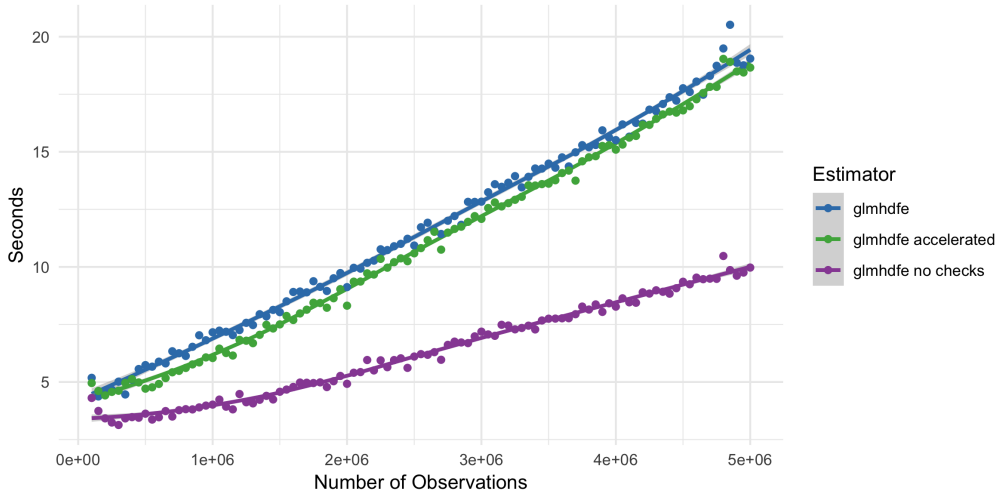

Figure 5: Time until convergence with fixed number of groups and increasing number of observations

References

Egger, P. H. and K. E. Staub (2016). GLM Estimation of Trade Gravity Models with Fixed Effects. *Empirical Economics 50*(1), 137–175.

Gaure, S. (2013). Ols with multiple high dimensional category variables. *Computational Statistics & Data Analysis 66*, 8–18.

Head, K. and T. Mayer (2014). Gravity Equations: Workhorse, Toolkit, and Cookbook. In G. Gopinath, E. Helpman, and K. Rogoff (Eds.), *Handbook of International Economics* (4 ed.), Volume 4, Chapter 3, pp. 131–195. North Holland.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (2 ed.). London/New York: Chapman and Hall.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Stammann, A. (2018). Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects. *arXiv preprint arXiv:1707.01815*.

A Link functions

A.1 Identity

With an identity link function, i.e. $g(\mu) = \mu$, we get the following FOCs:

$$\widehat{\boldsymbol{\beta}}: \quad \sum_i \frac{y_i - \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)}{V(y_i)}\boldsymbol{x}_i = \mathbf{0}, \tag{16a}$$

$$\widehat{\delta}^a: \quad \sum_{i, d_i^a = 1} \frac{y_i - \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)}{V(y_i)} = 0, \tag{16b}$$

$$\ldots$$

A.2 Log

Specifying an exponential conditional mean, i.e. the log-link function $g(\mu) = \log(\mu)$, yields the following general FOCs:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}: \quad & \sum_i \frac{y_i - \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)}{V(y_i)} \\ & \qquad \times \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)\boldsymbol{x}_i = \mathbf{0}, \end{aligned} \tag{17a}$$

$$\begin{aligned}\widehat{\delta}^a: \quad & \sum_{i, d_i^a = 1} \frac{y_i - \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right)}{V(y_i)} \\ & \qquad \times \exp\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{d}_i^A\right)'\widehat{\boldsymbol{\delta}}^A + \left(\boldsymbol{d}_i^B\right)'\widehat{\boldsymbol{\delta}}^B + \ldots\right) = 0, \end{aligned} \tag{17b}$$

$$\ldots$$