1 LMVAR: a linear model with heteroscedasticity

This vignette describes in more detail the mathematical aspects of the model with which the `lmvar` package is concerned. A short description can be found in the vignette 'Intro' of this package. The model has been discussed by various authors [1, 2, 3].

Assume that a stochastic vector $Y \in \mathbb{R}^n$ has a multivariate normal distribution as

$$Y \sim \mathcal{N}_n(\mu^\star, \Sigma^\star) \tag{1}$$

in which $\mu^\star \in \mathbb{R}^n$ is the expected value and $\Sigma^\star \in \mathbb{R}^{n,n}$ a diagonal covariance matrix

$$\Sigma^\star_{ij} = \begin{cases} 0 & i \neq j \\ (\sigma^\star_i)^2 & i = j. \end{cases} \tag{2}$$

Assume that the vector of expectation values $\mu^\star$ is linearly dependent on the values of the covariates in a model matrix X_μ:

$$\mu^\star = X_\mu \beta^\star_\mu \tag{3}$$

with $X_\mu \in \mathbb{R}^{n,k_\mu}$ and $\beta^\star_\mu \in \mathbb{R}^{k_\mu}$.

Similarly, assume that the vector $\sigma^\star = (\sigma^\star_1, \ldots, \sigma^\star_n)$ depends on the covariates in a model matrix X_σ as

$$\log \sigma^\star = X_\sigma \beta^\star_\sigma \tag{4}$$

where $\log \sigma^\star = (\log \sigma^\star_1, \ldots, \log \sigma^\star_n)$, $X_\sigma \in \mathbb{R}^{n,k_\sigma}$ and $\beta^\star_\sigma \in \mathbb{R}^{k_\sigma}$. The logarithm is taken to be the 'natural logarithm', i.e., with base e.

We assume $n \geq k_\mu + k_\sigma$ to avoid having an overdetermined system when we calculate estimators for $\beta^\star_\mu$ and $\beta^\star_\sigma$, as explained in the next section.

If we take X_σ a $n \times 1$ matrix in which each element is equal to 1, we have the standard linear model.

The parameter vector $\beta^\star_\mu$ is defined uniquely only if X_μ is full-rank. If not, the space $\mathbb{R}^{k_\mu}$ can be split into subspaces such that there is a uniquely defined $\beta^\star_\mu$ in each subspace. The way `lmvar` treats this is as follows. If the user-supplied X_μ is not full-rank, `lmvar` removes just enough columns from the matrix to make it full-rank. This amounts to selecting $\beta^\star_\mu$ from the subspace in which all vector elements corresponding to the removed columns, are set to zero.

In the same way, if the user-supplied X_σ is not full-rank, just enough columns are removed to make it so. This defines a subspace in which $\beta^\star_\sigma$ is defined uniquely.

In what follows we assume that X_μ and X_σ are the matrices after the columns have been removed, i.e., they are full-rank matrices. The vector elements that are set to zero, drop out of $\beta^\star_\mu$ and $\beta^\star_\sigma$ and the dimensions k_μ and k_σ are reduced accordingly. These reduced dimensions are returned by the function `dfree` in the `lmvar` package.

2 Maximum-likelihood equations

A vector element Y_i is distributed as

$$Y_i \sim \frac{1}{\sqrt{2\pi}\sigma_i^\star} \exp\left(-\frac{1}{2}\left(\frac{Y_i - \mu_i^\star}{\sigma_i^\star}\right)^2\right). \tag{5}$$

The logarithm of the likelihood $\mathcal{L}$ is defined as

$$\log \mathcal{L}(\beta_\mu, \beta_\sigma) = -\frac{n}{2}\log(2\pi) - \sum_{k=1}^{n}(\log \sigma_k + \frac{(y_k - \mu_k)^2}{2\sigma_k^2}). \tag{6}$$

for all vectors $\beta_\mu \in \mathbb{R}^{k_\mu}$ and $\beta_\sigma \in \mathbb{R}^{k_\sigma}$ and μ and σ defined as

$$\begin{aligned} \mu &= X_\mu \beta_\mu \\ \log \sigma &= X_\sigma \beta_\sigma. \end{aligned} \tag{7}$$

We are looking for $\hat{\beta_\mu} \in \mathbb{R}^{k_\mu}$ and $\hat{\beta_\sigma} \in \mathbb{R}^{k_\sigma}$ that maximize the log-likelihood:

$$(\hat{\beta}_\mu, \hat{\beta}_\sigma) = \operatorname*{argmax}_{(\beta_\mu, \beta_\sigma) \in \mathbb{R}^{k_\mu} \times \mathbb{R}^{k_\sigma}} \log \mathcal{L}(\beta_\mu, \beta_\sigma). \tag{8}$$

These maximum likelihood estimators are taken to be the estimators of $\beta_\mu^\star$ and $\beta_\sigma^\star$. We assume that $\hat{\beta}_\mu$ and $\hat{\beta}_\sigma$ thus defined, exist and are unique.

Given $\hat{\beta}_\sigma$, this is true for $\hat{\beta}_\mu$. Namely, given any β_σ, $\log \mathcal{L}$ is maximized by the β_μ which is the solution of

$$\nabla_{\beta_\mu} \log \mathcal{L} = 0 \tag{9}$$

where ∇_{β_μ} stands for the gradient $(\frac{\partial}{\partial \beta_{\mu,1}}, \ldots, \frac{\partial}{\partial \beta_{\mu,n}})$.

This solution is

$$\beta_\mu = \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \Sigma^{-1} y. \tag{10}$$

with $\Sigma \in \mathbb{R}^{n,n}$ defined as in (2) but with β_σ arbitrary:

$$\Sigma_{ij} = \begin{cases} 0 & i \neq j \\ \sigma_i^2 & i = j. \end{cases} \tag{11}$$

Because of our assumption that X_μ is full rank, the inverse of the matrix $X_\mu^T \Sigma^{-1} X_\mu$ can be taken.

It is easy to see that the solution (10) represents a maximum in the log-likelihood. The matrix $H_{\mu\mu}$ of second-order derivatives

$$(H_{\mu\mu})_{ij} = \frac{\partial^2 \log L}{\partial \beta_{\mu i} \partial \beta_{\mu j}} \tag{12}$$

is given by

$$H_{\mu\mu} = -X_\mu^T \Sigma^{-1} X_\mu, \tag{13}$$

which is negative-definite for any β_σ.

Our maximization search can now be carried out in a smaller space:

$$\hat{\beta_\sigma} = \underset{\beta_\sigma \in \mathbb{R}^{k_\sigma}}{\operatorname{argmax}} \ \log \mathcal{L}_P(\beta_\sigma) \tag{14}$$

where $\mathcal{L}_P$ is the so-called profile-likelihood

$$\mathcal{L}_P(\beta_\sigma) = \mathcal{L}(\beta_\mu(\beta_\sigma), \beta_\sigma). \tag{15}$$

with β_μ depending on β_σ as in (10).

To find $\hat{\beta}_\sigma$ from (14), we must solve

$$(\nabla_{\beta_\mu} \log \mathcal{L})\,(\nabla_{\beta_\sigma} \beta_\mu) + \nabla_{\beta_\sigma} \log \mathcal{L} = 0 \tag{16}$$

evaluated at $\beta_\mu = \beta_\mu(\beta_\sigma)$, and $(\nabla_{\beta_\sigma}\beta_\mu)$ the matrix

$$(\nabla_{\beta_\sigma}\beta_\mu)_{ij} = \frac{\partial \beta_{\mu i}}{\partial \beta_{\sigma j}}. \tag{17}$$

However, because of (9), the first term in (16) vanishes and we are left to solve

$$\nabla_{\beta_\sigma} \log \mathcal{L} = 0. \tag{18}$$

The derivatives that are the elements of this gradient are given by

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \beta_{\sigma i}} &= \sum_{k=1}^{n} (-(X_\sigma)_{ki} + \frac{(y_k - \mu_k)^2}{\sigma_k^2}(X_\sigma)_{ki}) \\
&= \sum_{k=1}^{n} (\frac{(y_k - \mu_k)^2}{\sigma_k^2} - 1)(X_\sigma)_{ki}.
\end{aligned} \tag{19}$$

The entire gradient can be written as a matrix-product as

$$\nabla_{\beta_\sigma} \log \mathcal{L} = X_\sigma^T \lambda_\sigma \tag{20}$$

with λ_σ a vector of length n whose elements $\lambda_{\sigma i}$ are

$$\lambda_{\sigma i} = \left(\frac{y_i - \mu_i}{\sigma_i}\right)^2 - 1. \tag{21}$$

The maximum-likelihood equations (18) take the form

$$X_\sigma^T \lambda_\sigma = 0. \tag{22}$$

The estimate μ of the expectation value that appears in λ_σ depends on β_σ as

$$\begin{aligned}
\mu &= X_\mu \beta_\mu \\
&= X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \Sigma^{-1} y \\
&= \Sigma^{-1/2} X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \Sigma^{-1/2} y
\end{aligned} \tag{23}$$

where the latter form is the more symmetric, with

$$\left(\Sigma^{-1/2}\right)_{ij} = \begin{cases} 0 & i \neq j \\ \dfrac{1}{\sigma_i} & i = j. \end{cases} \tag{24}$$

The vector $(y - \mu)/\sigma$, which i-th element is $(y_i - \mu_i)/\sigma_i$, can be written as

$$\frac{y-\mu}{\sigma} = \Sigma^{-1/2}\left[I - \Sigma^{-1/2} X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \Sigma^{-1/2}\right] y \tag{25}$$

in which $I \in \mathbb{R}^{n,n}$ is the identity matrix.

2.1 Profile-likelihood Hessian

Numerical procedures to solve the maximum-likelihood equations $X_\sigma^T \lambda_\sigma = 0$ involve the calculation of the Hessian H_P of the profile log-likelihood. H_P is the matrix of second-order derivatives of $\log \mathcal{L}_P$:

$$(H_P)_{ij} = \frac{\partial^2 \log \mathcal{L}_P}{\partial \beta_{\sigma j} \partial \beta_{\sigma i}} \tag{26}$$

Differentiation of (19) gives for the second-order derivatives

$$(H_P)_{ij} = -2 \sum_{k=1}^{n} (X_\sigma^T)_{ik} \frac{y_k - \mu_k}{\sigma_k^2} \left\{ \frac{\partial \mu_k}{\partial \beta_{\sigma j}} + (y_k - \mu_k)(X_\sigma)_{kj} \right\} \tag{27}$$

with $\partial\mu_k/(\partial\beta_{\sigma j})$ the element at row k and column j of the matrix $(\nabla_{\beta_\sigma}\mu)$. Given that $\mu = X_\mu \beta_\mu$ and β_μ is given by (10), the j-th column vector of the matrix is

$$\begin{aligned}
\frac{\partial \mu}{\partial \beta_{\sigma j}} &= X_\mu \frac{\partial \beta_\mu}{\partial \beta_{\sigma j}} \\
&= X_\mu \left\{ \frac{\partial \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1}}{\partial \beta_{\sigma j}} X_\mu^T \Sigma^{-1} + \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \frac{\partial \Sigma^{-1}}{\partial \beta_{\sigma j}} \right\} y \\
&= X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} \left\{ -X_\mu^T \frac{\partial \Sigma^{-1}}{\partial \beta_{\sigma j}} X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \Sigma^{-1} + X_\mu^T \frac{\partial \Sigma^{-1}}{\partial \beta_{\sigma j}} \right\} y \\
&= X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \frac{\partial \Sigma^{-1}}{\partial \beta_{\sigma j}} \left\{ -X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \Sigma^{-1} + I \right\} y \\
&= X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \frac{\partial \Sigma^{-1}}{\partial \beta_{\sigma j}} (y - \mu) \qquad (28)
\end{aligned}$$

The matrix $\partial\Sigma^{-1}/(\partial\beta_{\sigma j})$ takes the form

$$\begin{aligned}
\frac{\partial \Sigma^{-1}}{\partial \beta_{\sigma j}} &= \sum_{i=1}^{n} \frac{\partial \Sigma^{-1}}{\partial \sigma_i} \frac{\partial \sigma_i}{\partial \beta_{\sigma j}} \qquad (29) \\
&= -2 \begin{pmatrix} (X_\sigma)_{1j} & & 0 \\ & \ddots & \\ 0 & & (X_\sigma)_{nj} \end{pmatrix} \Sigma^{-1}
\end{aligned}$$

The j-th column vector of the matrix is

$$\frac{\partial \mu}{\partial \beta_{\sigma j}} = -2X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \begin{pmatrix} \frac{y_1-\mu_1}{\sigma_1^2} (X_\sigma)_{1j} \\ \vdots \\ \frac{y_n-\mu_n}{\sigma_n^2} (X_\sigma)_{nj} \end{pmatrix} \tag{30}$$

and the element $(\nabla_{\beta_\sigma}\mu)_{kj}$ of the matrix $(\nabla_{\beta_\sigma}\mu)$ is given by

$$\frac{\partial \mu_k}{\partial \beta_{\sigma j}} = -2 \sum_{l=1}^{n} \left(X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T\right)_{kl} \frac{y_l-\mu_l}{\sigma_l^2} (X_\sigma)_{lj}\,. \tag{31}$$

If we substitute this result in (27), we obtain for the element at row i and column j of the Hessian:

$$\begin{aligned} &(H_P)_{ij} = \\ &4 \sum_{k,l=1}^{n} (X_\sigma^T)_{ik} \frac{y_k-\mu_k}{\sigma_k^2} \left(X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T\right)_{kl} \frac{y_l-\mu_l}{\sigma_l^2} (X_\sigma)_{lj} + \\ &-2 \sum_{k=1}^{n} (X_\sigma^T)_{ik} \left(\frac{y_k-\mu_k}{\sigma_k}\right)^2 (X_\sigma)_{kj}. \end{aligned} \tag{32}$$

We can write the Hessian as a matrix-product as

$$H_P = X_\sigma^T \Lambda_1 X_\mu \left(X_\mu^T \Sigma^{-1} X_\mu\right)^{-1} X_\mu^T \Lambda_1 X_\sigma + X_\sigma^T \Lambda_2 X_\sigma \tag{33}$$

with two $n \times n$ diagonal matrices

$$(\Lambda_1)_{ij} = \begin{cases} 0 & i \neq j \\ 2\,\dfrac{y_i-\mu_i}{\sigma_i^2} & i = j \end{cases} \qquad (\Lambda_2)_{ij} = \begin{cases} 0 & i \neq j \\ -2\left(\dfrac{y_i-\mu_i}{\sigma_i}\right)^2 & i = j. \end{cases} \tag{34}$$

3 Distributions for estimators

Asymptotic theory of maximum-likelihood estimators tells that the vector of the combined estimators $(\hat\beta_\mu, \hat\beta_\sigma)$ as defined in (8), is distributed approximately as

$$(\hat\beta_\mu, \hat\beta_\sigma) \sim \mathcal{N}_{k_\mu+k_\sigma}\left((\beta_\mu^\star, \beta_\sigma^\star), \Sigma_{\beta\beta}\right) \qquad \text{for } n \text{ large.} \tag{35}$$

This distribution is valid in the limit of a large number of observations n.

The covariance matrix $\Sigma_{\beta\beta}$ is given in terms of the inverse Fisher information matrix I_n:

$$\Sigma_{\beta\beta} = \frac{1}{n} I_n^{-1}. \tag{36}$$

The Fisher information matrix is given in terms of the expected value of the Hessian at $\beta_\mu = \beta_\mu^\star$ and $\beta_\sigma = \beta_\sigma^\star$:

$$I_n = -\frac{1}{n} E[H^\star]. \tag{37}$$

The Hessian H is the Hessian of the full log-likelihood, in contrast to the profile-likelihood Hessian:

$$H^\star = \begin{pmatrix} H^\star_{\mu\mu} & H^\star_{\mu\sigma} \\ {H^\star_{\mu\sigma}}^T & H^\star_{\sigma\sigma} \end{pmatrix} \tag{38}$$

with the three block-matrices defined as

$$\left(H^\star_{\mu\mu}\right)_{ij} = \frac{\partial^2 \log L}{\partial\beta_{\mu i}\partial\beta_{\mu j}},\ \left(H^\star_{\mu\sigma}\right)_{ij} = \frac{\partial^2 \log L}{\partial\beta_{\mu i}\partial\beta_{\sigma j}},\ \left(H^\star_{\sigma\sigma}\right)_{ij} = \frac{\partial^2 \log L}{\partial\beta_{\sigma i}\partial\beta_{\sigma j}} \tag{39}$$

evaluated at $\beta_\mu = \beta^\star_\mu$ and $\beta_\sigma = \beta^\star_\sigma$.

We have already calculated $H_{\mu\mu}$ in (13). The other block matrices are given by

$$\begin{aligned}
\left(H^\star_{\mu\sigma}\right)_{ij} &= -2\sum_{k=1}^{n} \frac{y_k - \mu^\star_k}{{\sigma^\star_k}^2} \left(X_\mu\right)_{ki} \left(X_\sigma\right)_{kj} \\
\left(H^\star_{\sigma\sigma}\right)_{ij} &= -2\sum_{k=1}^{n} \left(\frac{y_k - \mu^\star_k}{\sigma^\star_k}\right)^2 \left(X_\sigma\right)_{ki} \left(X_\sigma\right)_{kj} .
\end{aligned}$$

In matrix notation:

$$H^\star_{\mu\mu} = -X_\mu^T {\Sigma^\star}^{-1} X_\mu, \qquad H^\star_{\mu\sigma} = -X_\mu^T \Lambda^\star_1 X_\sigma, \qquad H^\star_{\sigma\sigma} = X_\sigma^T \Lambda^\star_2 X_\sigma. \tag{40}$$

with $\Lambda^\star_1$ equal to Λ_1 with $\mu = \mu^\star$ and $\sigma = \sigma^\star$, and likewise for $\Lambda^\star_2$.

When we take expected values and keep in mind that

$$\begin{aligned}
E[Y - \mu^\star] &= 0 \\
E[(Y_i - \mu^\star_i)(Y_j - \mu^\star_j)] &= \begin{cases} 0 & i \neq j \\ {\sigma^\star_i}^2 & i = j \end{cases},
\end{aligned}$$

we arrive at

$$E[H^\star_{\mu\mu}] = -X_\mu^T {\Sigma^\star}^{-1} X_\mu,\ E[H^\star_{\mu\sigma}] = 0,\ E[H^\star_{\sigma\sigma}] = -2X_\sigma^T X_\sigma \tag{41}$$

This brings the expected value of the Hessian in the form

$$E[H^\star] = -\begin{pmatrix} X_\mu^T {\Sigma^\star}^{-1} X_\mu & 0 \\ 0 & 2X_\sigma^T X_\sigma \end{pmatrix}. \tag{42}$$

The function `fisher` in the `lmvar` package calculates the Fisher information matrix. It estimates $E[H^\star]$ by replacing the true but unknown $\sigma^\star$ by its maximum-likelihood estimator $\hat{\sigma}$ in $\Sigma^\star$.

The expectation value (42) brings the covariance matrix $\Sigma_{\beta\beta}$ in the form

$$\Sigma_{\beta\beta} = \begin{pmatrix} \left(X_\mu^T {\Sigma^\star}^{-1} X_\mu\right)^{-1} & 0 \\ 0 & \frac{1}{2}\left(X_\sigma^T X_\sigma\right)^{-1} \end{pmatrix}. \tag{43}$$

This implies that $\hat{\beta}_\mu$ and $\hat{\beta}_\sigma$ are independent stochastic variables distributed as

$$\begin{aligned} \hat{\beta}_\mu &\sim \mathcal{N}_{k_\mu}(\beta^\star_\mu,\, \left(X_\mu^T \Sigma^{\star -1} X_\mu\right)^{-1}) \\ \hat{\beta}_\sigma &\sim \mathcal{N}_{k_\sigma}(\beta^\star_\sigma,\, \tfrac{1}{2}\left(X_\sigma^T X_\sigma\right)^{-1}) \end{aligned} \qquad \text{for } n \text{ large.} \tag{44}$$

We obtain for the asymptotic distribution of the maximum-likelihood estimators of $\mu^\star$ and $\sigma^\star$

$$\begin{aligned} \hat{\mu} &\sim \mathcal{N}_n(\mu^\star,\, X_\mu \left(X_\mu^T \Sigma^{\star -1} X_\mu\right)^{-1} X_\mu^T) \\ \log \hat{\sigma} &\sim \mathcal{N}_n(\log \sigma^\star,\, \tfrac{1}{2} X_\sigma \left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T) \end{aligned} \qquad \text{for } n \text{ large.} \tag{45}$$

The expectation value and the variance for an element $\hat{\sigma}_i$ of $\hat{\sigma}$ are

$$\begin{aligned} E[\hat{\sigma}_i] &= \sigma_i^\star \exp\left(\frac{(X_\sigma \left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T)_{ii}}{4}\right) \\ \operatorname{var}(\hat{\sigma}_i) &= \left(E[\hat{\sigma}_i]\right)^2 \left(\exp(\frac{(X_\sigma \left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T)_{ii}}{2}) - 1\right) \end{aligned} \qquad \text{for } n \text{ large.} \tag{46}$$

The function `fitted.lmvar` (with the option `log = FALSE`) returns $\hat{\mu}$ and $\hat{\sigma}$.

References

[1] Murray Aitkin. Modelling Variance Heterogeneity in Normal Regression Using GLIM. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):332–339, 1987.

[2] A. C. Harvey. Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica*, 44(3):461–465, 1976.

[3] A. P. Verbyla. Modelling Variance Heterogeneity: Residual Maximum Likelihood and Diagnostics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(2):493–508, 1993.