

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

Bayesian Conditional Mean Estimation in Log-Normal Linear Regression Models with Finite Quadratic Expected Loss

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Enrico, F., Trivisano, C. (2016). Bayesian Conditional Mean Estimation in Log-Normal Linear Regression Models with Finite Quadratic Expected Loss. SCANDINAVIAN JOURNAL OF STATISTICS, 43(4), 1064-1077 [10.1111/sjos.12229].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/568956> since: 2016-11-21

*Published:*

DOI: <http://doi.org/10.1111/sjos.12229>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

**Fabrizi, E., and Trivisano, C. (2016) Bayesian Conditional Mean Estimation in Log-Normal Linear Regression Models with Finite Quadratic Expected Loss. Scand J Statist, 43: 1064– 1077. doi: 10.1111/sjos.12229.**

The final published version is available online at: <https://doi.org/10.1111/sjos.12229>

#### Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)*

***When citing, please refer to the published version.***

# Bayesian conditional mean estimation in log-normal linear regression models with finite quadratic expected loss

ENRICO FABRIZI

*DISES, Università Cattolica, Piacenza*

CARLO TRIVISANO

*DS, Università di Bologna*

## Abstract

Log-normal linear regression models are popular in many fields of research. Bayesian estimation of the conditional mean of the dependent variable is problematic as many choices of the prior for the variance (on the log-scale) lead to posterior distributions with no finite moments. We propose a generalized inverse gaussian prior for this variance and derive the conditions on the prior parameters that yield posterior distributions of the conditional mean of the dependent variable with finite moments up to a pre-specified order. The conditions depend on one of the three parameters of the suggested prior; the other two have an influence on inferences for small and medium sample sizes. A second goal of this paper is to discuss how to choose these parameters according to different criteria including the optimization of frequentist properties of posterior means.

**Keywords:** Generalized inverse Gaussian, Generalized hyperbolic distribution, efficient estimation, prior specification

## 1 Introduction

Log-normal linear regression models are common in many fields of applied research including environmental sciences (El-Shaarawi and Viveros, 1997), medicine (Olin et al., 2007; Ahn et al., 2006), economics (Zellner et al., 1966; Mankiw et al., 1992), life testing and reliability studies (Nelson, 1990). In its simplest form, that we consider in this paper, the log-normal regression model may be described as follows. Let  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , be a random sample,  $y$  the dependent variable and  $\mathbf{x}_i$  a  $(p + 1) \times 1$  vector of covariates. We assume that

$$y_i \sim \text{Logn}(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \tag{1}$$

that implies a linear specification for the conditional expectation of the logarithmic transformation of the dependent variable  $z = \log(y)$ , that is  $z_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$  or equivalently  $z_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i$  where  $u_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$ .

Given a new point  $\mathbf{x}_0$  in the covariates' space, a researcher may be interested in estimating the

conditional mean of the response variable:

$$\theta(\mathbf{x}_0) = E(y_0|\mathbf{x}_0) = \exp\left(\mathbf{x}_0^T\boldsymbol{\beta} + \frac{1}{2}\sigma^2\right) \quad (2)$$

This estimation problem has been considered by many authors in the literature (Finney, 1941; Bradu and Mundlak, 1970; El-Shaarawi and Viveros, 1997; Shen and Zhu, 2008).

If, in line with many applications, we assume a flat prior on the regression coefficients, i.e.  $p(\boldsymbol{\beta}) \propto 1$ , it easily follows that  $\mathbf{x}_0^T\boldsymbol{\beta}|\mathbf{y}, \sigma^2 \sim N(\mathbf{x}_0^T\hat{\boldsymbol{\beta}}, h_{00}\sigma^2)$  where  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}$ ,  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ ,  $\mathbf{z} = (z_1, \dots, z_n)$ ,  $\mathbf{y} = (y_1, \dots, y_n)$  and  $h_{00} = \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0$ . Similarly, for  $\eta(\mathbf{x}_0) = \log\{\theta(\mathbf{x}_0)\}$  we have that  $\eta(\mathbf{x}_0)|\mathbf{y}, \sigma^2 \sim N(\mathbf{x}_0^T\hat{\boldsymbol{\beta}} + \frac{1}{2}\sigma^2, h_{00}\sigma^2)$ , so it follows that  $\theta(\mathbf{x}_0)|\mathbf{y}, \sigma^2 \sim \text{Logn}(\mathbf{x}_0^T\hat{\boldsymbol{\beta}} + \frac{1}{2}\sigma^2, h_{00}\sigma^2)$  a posterior distribution that may be summarized using a quadratic loss function or a relative quadratic loss function, as suggested by Zellner (1971), in order to obtain a point predictor with smaller frequentist MSE. A problem closely related to that of estimating (2) is the prediction of  $y_0 = \exp(\mathbf{x}_0\boldsymbol{\beta} + u_0)$  that is relevant for instance in finite population inference (Karlberg, 2000). We note that since  $\mathbf{x}_0^t\boldsymbol{\beta}|\mathbf{y}, \sigma^2 \sim N(\mathbf{x}_0^t\hat{\boldsymbol{\beta}}, h_{00}\sigma^2)$  then  $\mathbf{x}_0^t\boldsymbol{\beta} + \varepsilon_0|\mathbf{y}, \sigma^2 \sim N(\mathbf{x}_0^t\hat{\boldsymbol{\beta}}, (1 + h_{00})\sigma^2)$ . For this reason, our results on the estimation of (2) readily extends to the problem of predicting  $y_0$ .

If a reference prior  $p(\sigma^2) \propto \sigma^{-2}$  is assumed for the residuals' variance we obtain that  $\mathbf{x}_0^T\boldsymbol{\beta}|\mathbf{y} \sim t(n - p, \mathbf{x}_0\hat{\boldsymbol{\beta}}, S^2h_{00})$  where  $S^2 = \sum_{i=1}^n (z_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2 / (n - p - 1)$ . As a consequence

$$\exp\{\mathbf{x}_0^T\boldsymbol{\beta}\}|\mathbf{y} \sim \text{logt}(n - p - 1, \mathbf{x}_0\hat{\boldsymbol{\beta}}, S^2h_{00}). \quad (3)$$

Moments of all orders of the *logt* distribution do not exist, making impossible to summarize this posterior using popular loss functions, such as the quadratic. Similarly, moments of  $\theta(\mathbf{x}_0)|\mathbf{y}$  do not exist as well, as noted also by Zellner (1971, footnote n. 9). This problem, that we illustrated for the reference prior  $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$ , remains for other popular prior choices including informative normal priors for  $\boldsymbol{\beta}$ , Jeffrey's, Inverse Gamma prior for  $\sigma^2$ , as it will be shown later on. Moreover it is not specific to Bayesian analysis of the model. We note that the marginal sampling distribution of  $\hat{z}_0 = \mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ , the ordinary least square predictor is *t* so its exponential transform will be characterized by non-existing sampling moments.

In this article we suggest the use of a generalized inverse Gaussian prior for  $\sigma^2$ . This rich family of distributions includes gamma, inverse gamma, inverse Gaussian and many others as special cases. In the analysis of the regression model it has been considered by several authors

(Tabane and Haq, 2008; Griffin et al., 2010; Li and Lin, 2010), while Fabrizi and Trivisano (2012) use it in the analysis of the special case of (1) obtained  $\mathbf{x}_i^T \boldsymbol{\beta} = \xi$ . We derive the conditions on its parameters that guarantee the existence of posterior moments of order  $r$ , ( $r > 0$ ) for  $\theta(\mathbf{x}_0)$ . This result extends the one in Fabrizi and Trivisano (2012) to the multivariate regression setting.

The choice of prior parameters may have a considerable impact on the posterior distribution of quantities of interest in the analysis of the log-normal linear model for small and moderate sample sizes. A second objective of this paper is to discuss how these parameters can be chosen to achieve specific inferential goals. A first choice is based on the idea of resuming posterior mean and variance we would get for  $p(\boldsymbol{\beta}|\mathbf{y})$  under a reference prior  $p(\sigma^2) \propto \sigma^{-2}$ . We consider also choice criteria based on minimization of frequentist mean square error (MSE) of  $\hat{\theta}_B(\mathbf{x}_0) = E\{\theta(\mathbf{x}_0)|\mathbf{y}\}$ , the ‘Bayes estimator’ of  $\theta(\mathbf{x}_0)$  under quadratic loss. This choice leads to estimators that compare favourably with known alternatives in the literature (Shen and Zhu, 2008) in terms of frequentist MSE.

The paper is organized as follows. In section 2, we shortly review the generalized inverse Gaussian and the generalized hyperbolic distributions; the latter is involved as  $p\{\eta(\mathbf{x}_0)|\mathbf{y}\}$  is within this class. In section 3, we present the analysis of the log-normal regression model under the proposed prior distribution for  $\sigma^2$ . Section 4 is about prior parameters choice. In section 5,  $\hat{\theta}_B(\mathbf{x}_0)$  is compared to alternative estimators of  $\theta(\mathbf{x}_0)$  in terms of frequentist properties using a simulation exercise similar to that considered in Shen and Zhu (2008). Section 6 considers an empirical application from the field of accelerated life testing. Section 7 offers some conclusions. The proofs of the theorems (Appendix A) and additional results from the simulation exercise (Appendix B) are available in the on-line supporting information file.

## 2 The generalized inverse Gaussian and generalized hyperbolic distributions

In this section we introduce the generalized inverse Gaussian (GIG) and generalized hyperbolic (GH) distributions, establish the notation and list some properties that will be used in subsequent sections. For more details on these distributions, see Bibby and Sørensen (2003) and Eberlein and von Hammerstein (2004), among others.

The density of the GIG distribution may be written as follows:

$$p(x) = \left(\frac{\gamma}{\delta}\right)^\lambda \frac{1}{2K_\lambda(\delta\gamma)} x^{\lambda-1} \exp\left\{-\frac{1}{2}(\delta^2 x^{-1} + \gamma^2 x)\right\} \mathbf{1}_{\mathbb{R}^+} \quad (4)$$

The permitted parameters are  $\delta > 0$ ,  $\gamma \geq 0$  if  $\lambda < 0$ ;  $\delta, \gamma < 0$  if  $\lambda = 0$ ;  $\delta \geq 0$ ,  $\gamma > 0$  if  $\lambda > 0$ .

The moments of the GIG can be expressed as functions of the modified Bessel functions of the third kind (the Bessel-K function from now on):

$$E(X^j) = \left(\frac{\delta}{\gamma}\right)^j \frac{K_{\lambda+j}(\delta\gamma)}{K_\lambda(\delta\gamma)}. \quad (5)$$

Fabrizi and Trivisano (2012) show that, for  $\gamma > 0$ , the approximation

$$E(X) \cong \frac{\lambda + \sqrt{(\lambda^2 + \delta^2 \gamma^2)}}{\gamma^2} \quad (6)$$

is adequate in most cases. Many important distributions may be obtained as special cases of the GIG. For  $\lambda > 0$  and  $\gamma > 0$ , the gamma distribution emerges as the limit when  $\delta \rightarrow 0$ . The inverse-gamma is obtained when  $\lambda < 0$ ,  $\delta > 0$  and  $\gamma \rightarrow 0$  and an inverse Gaussian distribution is obtained when  $\lambda = -\frac{1}{2}$ .

Barndorff-Nielsen (1977) introduce a general class of multivariate distributions, the multivariate generalized hyperbolic (MVGH) distributions as scale mixtures of Gaussian vectors. Specifically, if  $(\mathbf{X}|W = w) \sim MVN_d(\boldsymbol{\mu} + w\boldsymbol{\psi}\Delta, w\Delta)$  and  $W \sim GIG(\lambda, \delta, \sqrt{\alpha^2 - \boldsymbol{\psi}^T \Delta \boldsymbol{\psi}})$  then  $\mathbf{X} \sim MVGH_d(\lambda, \alpha, \boldsymbol{\mu}, \Delta, \boldsymbol{\psi})$  which is characterized by the density

$$f(\mathbf{x}) = \frac{(\alpha^2 - \boldsymbol{\psi}^T \Delta \boldsymbol{\psi})^{\lambda/2}}{(2\pi)^{d/2} \sqrt{|\Delta|} \alpha^{\lambda-1/2}} \frac{K_{\lambda-\frac{d}{2}}(\alpha \sqrt{\delta^2 + (\mathbf{x} - \boldsymbol{\mu})^T \Delta^{-1} (\mathbf{x} - \boldsymbol{\mu})}) e^{-\boldsymbol{\psi}^T (\mathbf{x} - \boldsymbol{\mu})}}{K_\lambda(\delta \sqrt{\alpha^2 - \boldsymbol{\psi}^T \Delta \boldsymbol{\psi}}) (\sqrt{\delta^2 + (\mathbf{x} - \boldsymbol{\mu})^T \Delta^{-1} (\mathbf{x} - \boldsymbol{\mu})})^{\frac{d}{2}-\lambda}} \quad (7)$$

The parameters domain is:  $\lambda \in \mathbb{R}$ ,  $\alpha > 0$ ,  $\delta \geq 0$ ,  $\boldsymbol{\psi} \in \{\mathbf{u} \in \mathbb{R}^d : \alpha^2 - \mathbf{u}^T \Delta \mathbf{u} > 0\}$ ;  $\Delta$  is a semi-positive definite matrix. Sometimes the constraint  $|\Delta| = 1$  is added to solve identifiability problems in estimation but it is unnecessary here. We have that:  $E(\mathbf{X}) = \boldsymbol{\mu} + E(W)\Delta\boldsymbol{\psi}$  and  $V(\mathbf{X}) = V(W)\Delta\boldsymbol{\psi}\boldsymbol{\psi}^T \Delta + E(W)\Delta$ . More details about this distribution can be found in Breyman and Lüthi (2013).

### 3 Analysis of the log-normal regression model using generalized inverse Gaussian prior for $\sigma^2$

As anticipated in the Introduction, let  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , be a random sample  $y$  the dependent variable and  $\mathbf{x}_i$  a  $p \times 1$  vector of covariates. Let's assume that  $y_i \sim \text{Logn}(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$  or equivalently, once defined  $z_i = \log(y_i)$ ,  $z_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$ . We are interested in making inference on the model parameters  $(\boldsymbol{\beta}, \sigma^2)$  and the conditional mean  $\theta(\mathbf{x}_0)$  (see 2). We assume the following priors for  $(\boldsymbol{\beta}, \sigma^2)$ :

$$p(\boldsymbol{\beta}|\sigma^2) = MVN_{p+1}(\boldsymbol{\beta}_0, \sigma^2 \mathbf{V}_0) \quad (8)$$

where  $\boldsymbol{\beta}_0$  is  $(p+1) \times 1$  vector and  $\mathbf{V}_0$  a  $(p+1) \times (p+1)$  positive definite matrix of known constants. The popular Zellner's  $g$ -priors are within this class provided that we set  $\mathbf{V}_0 = g(\mathbf{X}^T \mathbf{X})^{-1}$ . The smaller  $g$  is, the more weight the prior receives, while the case of non informative flat prior on  $\boldsymbol{\beta}$  is recovered for  $g \rightarrow +\infty$ . As for  $\sigma^2$  we assume:

$$p(\sigma^2) = GIG(\lambda, \delta, \gamma) \quad (9)$$

which encompasses many important special cases as explained in section 2.

Conditionally on  $\sigma^2$ , the following results follow from standard Bayesian analysis of the linear model: *i*)  $p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) = MVN_{p+1}(\boldsymbol{\beta}_*, \mathbf{V}_*)$  where  $\boldsymbol{\beta}_* = \mathbf{V}_*(\mathbf{X}^T \mathbf{y} + \mathbf{V}_0^{-1} \boldsymbol{\beta}_0)$  and  $\mathbf{V}_* = (\mathbf{X}^T \mathbf{X} + \mathbf{V}_0^{-1})^{-1}$ ; *ii*)  $p(\mathbf{x}_0^T \boldsymbol{\beta}|\sigma^2, \mathbf{y}) = N(\mathbf{x}_0^T \boldsymbol{\beta}_*, \sigma^2 h_{0*})$ ; with  $h_{0*} = \mathbf{x}_0^T \mathbf{V}_* \mathbf{x}_0$ ; *iii*)  $p\{\eta(\mathbf{x}_0)|\sigma^2, \mathbf{y}\} = N(\mathbf{x}_0^T \boldsymbol{\beta}_* + \frac{1}{2} \sigma^2, \sigma^2 h_{0*})$ . The main results on the posterior distributions obtained integrating out  $\sigma^2$  are summarized in the following

**Theorem 3.1.** *If the log-normal regression model (1) is assumed and (8), (9) are specified as priors for  $\boldsymbol{\beta}$  and  $\sigma^2$  then:*

$$i) \quad p(\sigma^2|\mathbf{y}) = GIG(\bar{\lambda}, \bar{\delta}_*, \gamma) \quad (10)$$

where  $\bar{\lambda} = \lambda - \frac{n-p-1}{2}$ ,  $\bar{\delta}_* = \sqrt{RSS + \delta^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathbf{X}^T \mathbf{X} \mathbf{V}_* \mathbf{V}_0^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}$  and  $RSS = (n - p - 1)S^2$ ;

$$ii) \quad p(\boldsymbol{\beta}|\mathbf{y}) = MVGH_{p+1}(\bar{\lambda}, \gamma, \boldsymbol{\beta}_*, \mathbf{V}_*, \bar{\delta}_*, 0) \quad (11)$$

$$iii) \quad p\{\eta(\mathbf{x}_0)|\mathbf{y}\} = GH(\bar{\lambda}, \bar{\alpha}, \bar{\mu}, \bar{\delta}, \bar{\psi}) \quad (12)$$

where  $\bar{\alpha} = \frac{\sqrt{h_{0\star}\gamma^2 + \frac{1}{4}}}{h_{0\star}}$ ,  $h_{0\star} = \mathbf{x}_0^T \mathbf{V}_\star \mathbf{x}_0$ ,  $\bar{\mu} = \mathbf{x}_0^T \boldsymbol{\beta}_\star$ ,  $\bar{\delta} = \sqrt{h_{0\star}\delta_\star}$ ,  $\bar{\psi} = (2h_{0\star})^{-1}$ . Moreover  $\bar{\gamma}^2 = \bar{\alpha}^2 - \bar{\psi}^2 = \gamma^2/h_{0\star}$ .

*Proof.* See appendix A in the online Supporting Information.  $\square$

Formula (10) states that the GIG prior on  $\sigma^2$  is conjugate to the normal likelihood and leads to a very easily manageable posterior distribution. In particular, assuming  $\bar{\lambda} < 0$ , using (6) and a first order expansion of the square root around  $\lambda$  we have that  $E(\sigma^2|\mathbf{y}) \cong \bar{\delta}_\star^2(2|\bar{\lambda}|)^{-1}$ . Using formulas for the moments of the multivariate generalized hyperbolic distribution we get that  $E(\boldsymbol{\beta}|\mathbf{y}) = \boldsymbol{\beta}_\star$  and approximately that  $V(\boldsymbol{\beta}|\mathbf{y}) = \mathbf{V}_\star \frac{\bar{\delta}_\star^2}{2|\bar{\lambda}|}$ .

If we assume an improper flat prior on  $\boldsymbol{\beta}$  (i.e. we set  $\mathbf{V}_0 = k\mathbf{I}_{p+1}$  and let  $k \rightarrow +\infty$ ) we have that  $\bar{\delta}_\star^2 = \sqrt{RSS + \delta^2}$  and

$$E(\sigma^2|\mathbf{y}) \cong \frac{RSS + \delta^2}{n - p - 1 - 2\lambda} \quad (13)$$

For the same prior choice, we have that  $\boldsymbol{\beta}_\star = \hat{\boldsymbol{\beta}}$ ,  $\mathbf{V}_\star = (\mathbf{X}^T \mathbf{X})^{-1}$  and in view of (13) we obtain

$$V(\boldsymbol{\beta}|\mathbf{y}) \cong (\mathbf{X}^T \mathbf{X})^{-1} \frac{RSS + \delta^2}{n - p - 1 - 2\lambda} \quad (14)$$

From (12) it follows that the posterior distribution of  $\theta(\mathbf{x}_0)$  will be a log-generalized hyperbolic distribution, whose moment of order  $r$  exists whenever the moment generating function is defined for that  $r$ . We can state the following

**Theorem 3.2.** *If the log-normal regression model (1) is assumed and (8), (9) are specified as priors for  $\boldsymbol{\beta}$  and  $\sigma^2$ , we have that  $E\{\theta(\mathbf{x}_0)|\mathbf{y}\}^r < +\infty$  if and only if the parameter  $\gamma$  in (9) is such that:*

$$\gamma^2 > h_{0\star} \left( r^2 + \frac{r}{h_{0\star}} \right) \quad (15)$$

*Proof.* See appendix A in the online Supporting Information.  $\square$

From theorem 3.2 we get that the conditions on the parameters of the GIG prior for  $\sigma^2$  that guarantee  $E\{\theta(\mathbf{x}_0)|\mathbf{y}\}^r < +\infty$  reduce to a single condition on the  $\gamma$  parameter. In fact, the parameter  $\gamma$  rules the weight of the right tail of the GIG distribution. Although we keep a  $(0, +\infty)$  support for the prior of  $\sigma^2$ , the idea is that of specifying a prior with a very light tail. In fact the smaller  $\gamma$ , the heavier the right tail of the GIG is. The popular inverse-gamma prior is a limit case as it can be obtained as a special case of the GIG for  $\lambda < 0$ ,  $\delta > 0$  and  $\gamma \rightarrow 0$ . Thereby this choice leads to  $E\{\theta(\mathbf{x}_0)|\mathbf{y}\}^r = +\infty \forall r$ . The same happens to the reference prior mentioned in the Introduction that may be approximated by  $p(\sigma^2) = \text{Invgamma}(\varepsilon, \varepsilon)$  for some ‘small’  $\varepsilon$ . Similarly,



the uniform prior over the range  $(0, A)$  for  $\sigma$  (Gelman, 2006) implies that  $p(\sigma^2) \propto \frac{1}{\sigma} \mathbf{1}_{(0,A)}$ , which may be seen as an approximation to a  $Gamma(\frac{1}{2}, \epsilon)$  (where  $\epsilon = (4A^2)^{-1}$ ) truncated at  $A^2$ . For  $\lambda > 0$ ,  $\gamma > 0$  and  $\delta \rightarrow 0$ ,  $GIG(\lambda, \delta, \gamma) \rightarrow Gamma(\lambda, \gamma^2/2)$ . If we let  $A \rightarrow \infty$ , therefore,  $p(\sigma) \propto 1$  is equivalent to a GIG prior with  $\gamma \rightarrow 0$  and thus implies non-existent posterior moments.

Condition (15) depends also on  $h_{0\star}$ . Suppose we adopt a Zellner's g-prior ( $\mathbf{V}_0 = g(\mathbf{X}^T \mathbf{X})^{-1}$ ) then  $h_{0\star}$  reduces  $h_{0g} = \frac{g}{g+1} h_{00}$  and (15) to  $\gamma^2 > \frac{g}{g+1} h_{00} r^2 + r$ . The more information we incorporate in  $p(\beta)$ , i.e. the smaller  $g$ , the milder will be condition (15), even though we need  $\gamma^2 > r$  even when  $g \rightarrow 0$ .

## 4 Choice of the generalized inverse Gaussian prior parameters

Posterior distributions derived in theorem 3.1 are sensitive to the choice of prior parameters  $(\beta_0, \mathbf{V}_0)$  and  $(\lambda, \delta, \gamma)$ . As the impact of prior information on  $\beta$  on related posteriors has been widely studied in the literature, in this section we focus on the choice of the GIG prior parameters assuming  $\beta_0 = 0$  and  $\mathbf{V}_0 = k\mathbf{I}_{p+1}$ ,  $k \rightarrow +\infty$ .

We discuss three different choices of the GIG prior parameters. The first choice represents a default or ‘reference’ option that leads to posterior mean and variance of  $p(\beta|\mathbf{y})$  matching the least square frequentist estimator of  $\beta$  and its variance (approximately in this latter case). The remaining choices are based on the minimization of the frequentist mean square error of the Bayes estimator  $\hat{\theta}_B(\mathbf{x}_0)$ . The resulting priors are  $\mathbf{x}_0$ -specific. The second choice of priors parameters we suggest, eliminates this dependence using average values or other conservative elicitations, while the third exploits this prediction specificity to obtain a Bayes estimator which is very competitive in terms of efficiency with respect to other frequentist alternatives. In all cases  $\gamma$  will be chosen to be big enough to guarantee finite quadratic loss for the estimation of the conditional mean  $\theta(\mathbf{x}_0)$ , i.e. finite posterior expectation and variance.

### 4.1 Reference choice of prior parameters

We propose to choose  $\lambda = 0$ ,  $\delta = \varepsilon$  for some ‘small’  $\varepsilon$  (such as  $\varepsilon = 0.01$  or  $\varepsilon = 0.001$ ) and

$$\gamma = \gamma_m = \sqrt{m(4 + m^{-1})} + \varepsilon \quad (16)$$

where  $m = \max\{h_{ii}\}$  and  $h_{ii} = \mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ ,  $i$  ranging over sample observations. We note that  $m < 1$  but it is much smaller than 1 in most practical situations.

The choice of  $\lambda = 0$ ,  $\delta = \varepsilon$  leads to  $E(\sigma^2|\mathbf{y}) = S^2$  according to (13) and  $V(\boldsymbol{\beta}|\mathbf{y}) \cong (\mathbf{X}^T\mathbf{X})^{-1}S^2$  according to (14), that is those obtained under the same flat improper prior for  $\boldsymbol{\beta}$  and the reference prior  $p(\sigma^2) \propto \sigma^{-2}$  mentioned in the introduction. Moreover, since  $E\{\eta(\mathbf{x}_0)|\mathbf{y}\} = \mathbf{x}_0^T\hat{\boldsymbol{\beta}} + \bar{\psi}E(h_{00}\sigma^2|\mathbf{y})$  and  $h_{00}\sigma^2|\mathbf{y} \sim GIG(\bar{\lambda}, \sqrt{h_{00}}\bar{\delta}_0, \gamma/\sqrt{h_{00}})$  (with  $\bar{\delta}_0 = \sqrt{RSS}$ ) then  $E\{\eta(\mathbf{x}_0)|\mathbf{y}\} \cong \mathbf{x}_0^T\hat{\boldsymbol{\beta}} + S^2/2$  which is a frequentistically unbiased estimator of  $\eta(\mathbf{x}_0)$ , a property that is helpful when building posterior probability intervals with good frequentist properties.

The suggested choice for  $\gamma$  is aimed at getting finite expectations and variances for  $p\{\theta(\mathbf{x}_0)|\mathbf{y}\}$  for any  $h_{00} < m$ , a conservative choice that may be modified for specific prediction purposes.

## 4.2 Choice of prior parameters based on frequentist optimality criteria

In this section we discuss how the parameters of  $p(\sigma^2) = GIG(\lambda, \delta, \gamma)$  can be chosen in order to minimize the frequentist mean square error ( $MSE_s$ ) of

$$\begin{aligned}\hat{\theta}^B(\mathbf{x}_0) &= \exp(\bar{\mu}) \left\{ \frac{\bar{\gamma}^2}{\bar{\alpha}^2 - (\bar{\psi} + 1)^2} \right\}^{\bar{\lambda}/2} \frac{K_{\bar{\lambda}}(\bar{\delta}\sqrt{\bar{\alpha}^2 - (\bar{\psi} + 1)^2})}{K_{\bar{\lambda}}(\bar{\delta}\bar{\gamma})} \\ &= \exp(\mathbf{x}_0^T\hat{\boldsymbol{\beta}}) \left\{ \frac{\gamma^2}{\gamma^2 - (h_{00} + 1)} \right\}^{\left(\frac{2\lambda - n - p - 1}{4}\right)} \frac{K_{\{\lambda - \frac{n-p-1}{2}\}}\left(\sqrt{(RSS + \delta^2)\{\gamma^2 - (h_{00} + 1)\}}\right)}{K_{\{\lambda - \frac{n-p-1}{2}\}}\left(\sqrt{(RSS + \delta^2)\gamma^2}\right)}.\end{aligned}$$

This expression is obtained using the moment generating function of the univariate generalized hyperbolic distribution; in view of (15)  $\hat{\theta}^B(\mathbf{x}_0)$  is defined only if  $\gamma^2 > h_{00} + 1$ .

We look for the  $MSE_s$ -minimizing prior parameters first restricting the class of Bayes estimators to those in the form

$$\tilde{\theta}(\mathbf{x}_0) = \exp(\mathbf{x}_0^T\hat{\boldsymbol{\beta}})g(RSS) \quad (17)$$

with  $g(RSS) > 1$ . The starting point of our discussion is the following result that is in line with the approach of Rukhin (1986, section 2):

**Theorem 4.1.** *Given  $\tilde{\theta}(\mathbf{x}_0) = \exp(\mathbf{x}_0^T\hat{\boldsymbol{\beta}})g(RSS)$  with  $g(\cdot)$  a real function of  $Y$  such that  $g(RSS) > 1$ , then  $g(RSS)$  minimizing  $MSE_s\{\tilde{\theta}(\mathbf{x}_0)\}$  is the same that minimizes  $E_s\{g(RSS) - \exp(c_R\sigma^2)\}$  where*

$$c_R = \frac{1}{2}(1 - 3h_{00}) \quad (18)$$

*Proof.* See appendix A in the online Supporting Information.  $\square$

The presence in (17) of the Bessel-K functions makes the minimization of  $E_s\{g(RSS) - \exp(c_R\sigma^2)\}$  analytically untractable. To circumvent this problem we propose to use a ‘small argument’ approximation to the ratio of Bessel-K function in the (17), following an argument of Fabrizi and Trivisano (2012, theorem 4.1). This approximation leads to a much simpler expression.

**Theorem 4.2.** *Under the assumptions that i)  $(RSS + \delta^2)\gamma^2 < 1$ , ii)  $0 < (RSS + \delta^2)\{\gamma^2 - (h_{00} + 1)\} < 1$ , iii)  $\lambda < \frac{n-p-1}{2}$ , we have that:*

$$\hat{\theta}^B(\mathbf{x}_0) \cong \exp(\mathbf{x}_0^T \hat{\beta}) \exp\left\{\frac{(RSS + \delta^2)(h_{00} + 1)}{4\left(\frac{n-p-3}{2} - \lambda\right)}\right\} = \hat{\theta}^b(\mathbf{x}_0) \quad (19)$$

*Proof.* See appendix A in the online Supporting Information.  $\square$

Although conditions *i* and *ii* are quite restrictive, Fabrizi and Trivisano (2012) and also simulations applied in our context show that  $\hat{\theta}^b(\mathbf{x}_0)$  approximates  $\hat{\theta}^B(\mathbf{x}_0)$  adequately also when they are not satisfied. We note that (19) is free of  $\gamma$ . This is in line with the fact that  $\gamma$  is a shape parameter that rules the weight of the right tail of the GIG distribution; consistently (15) is expressed as a function of  $\gamma$  only. The main result of this section can now be stated.

**Theorem 4.3.** *Under the assumptions of theorem 4.2, the value of  $\lambda$  in (19) that minimizes  $E_s\{g(RSS) - \exp(c_R\sigma^2)\}$  is given by*

$$\lambda_{opt} = \frac{n-p-3}{2} - \frac{(h_{00} + 1)(n-p-1)}{4c_R} - \frac{(h_{00} + 1)\delta^2}{4c_R\sigma^2} \quad (20)$$

*Proof.* See appendix A in the online Supporting Information.  $\square$

We note that although  $g(RSS)$  in (19) is a function of  $(\lambda, \delta)$  the minimum of  $E_s\{g(RSS) - \exp(c_R\sigma^2)\}$  is not a single point but a set of pairs described by (20). We propose to choose  $\delta = \varepsilon$  with  $\varepsilon$  ‘small’ (e.g.  $\varepsilon = 0.01$  or  $\varepsilon = 0.001$ ), for the following reasons: *i*) a ‘small’  $\delta$  removes the dependence of  $\lambda_{opt}$  on the unknown  $\sigma^2$  so that

$$(\lambda_{opt}|\delta = \varepsilon) \cong \frac{n-p-3}{2} - \frac{(h_{00} + 1)(n-p-1)}{2(1-3h_{00})}; \quad (21)$$

*ii*) the choice of a ‘small’  $\delta$  is line with section 4.1. Moreover if we replace  $\lambda_{opt}$  calculated according to (21) and  $\delta = \varepsilon$  into (19) it easy to show that

$$\hat{\theta}^b(\mathbf{x}_0) \cong \exp\left\{\mathbf{x}_0^T \hat{\beta} + \frac{S^2}{2}(1-3h_{00})\right\} \quad (22)$$

that is exactly the same predictor suggested by Zellner (1971) under relative quadratic loss function, with the assumed known  $\sigma^2$  replaced by  $S^2$ . We note that  $\delta = 0$  is not viable: this choice would require a strictly positive  $\lambda$  which is inconsistent with (21) (see section 2 of this paper and Bibby and Sørensen, 2003, p. 213, for more details). From (22) it is apparent that  $h_{00} > \frac{1}{3}$  would lead to  $g(RSS) < 1$  and  $\lambda > \frac{n-p-1}{2}$ , a violation of the hypotheses of theorem 4.2. Moreover for  $h_{00} = \frac{1}{3}$   $\lambda_{opt}$  is not defined. For this reason, values of  $h_{00} > \frac{1}{3}$  cannot be considered in the hyperparameters choice strategy that we are illustrating.

As anticipated,  $\lambda_{opt}$  depends on  $\mathbf{x}_0$  through  $h_{00}$ . To avoid a prior that is prediction-specific we may replace  $h_{00}$  with  $\bar{h} = n^{-1}(p+1)$ . As  $\bar{h} = n^{-1} \sum_{i=1}^n h_{ii}$ , the average of the leverages associated to sample observations, this choice may heuristically be motivated as optimal with respect to a prediction associated to the unknown vector  $\mathbf{x}_*$  for which we assume a priori distribution that matches the empirical distribution of in-sample  $\mathbf{x}_i$ . Note that  $\bar{h} < \frac{1}{3}$  whenever  $n > 3(p+1)$ , a condition that is not very restrictive, even in small samples. According to this choice we obtain:

$$(\lambda_{opt} | \delta = \varepsilon, h_{00} = \bar{h}) \cong \frac{n-p-3}{2} - \frac{(n+p+1)(n-p-1)}{2\{n-3(p+1)\}} = \lambda_{opt1}. \quad (23)$$

If we are willing to accept a  $\mathbf{x}_0$ -specific prior, or considering  $\hat{\theta}^B(\mathbf{x}_0)$  as a frequentist estimator of  $\theta(\mathbf{x}_0)$ , we can set  $\lambda$  according to (21). To respect the constraint  $g(RSS) > 1$  and the assumptions underlying theorem 4.2 we introduce the truncated leverage  $h_{00}^{tr} = \min(h_{00}, \frac{1}{3} - \varepsilon)$  and consequently suggest:

$$(\lambda_{opt} | \delta = \varepsilon, h_{00} = h_{00}^{tr}) \cong \frac{n-p-3}{2} - \frac{(h_{00}^{tr} + 1)(n-p-1)}{2(1-3h_{00}^{tr})} = \lambda_{opt2}; \quad (24)$$

We note that the threshold of  $1/3$  is very large and seldom crossed in most applications. In case of  $p = 1$ , this crossing would imply that  $(\mathbf{x}_0 - \bar{\mathbf{x}})^2 > \frac{1}{3} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2$  (with  $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$ ). Moreover the constraint  $h_{00} < \frac{1}{3}$  is imposed only in the choice of the GIG prior parameters, with  $h_{00}$  appearing in (17) remains that specific to the prediction at hand.

## 5 A simulation exercise to assess the frequentist properties of $\hat{\theta}^B(\mathbf{x}_0)$

In this section we discuss a simulation exercise aimed at assessing the frequentist properties of  $\hat{\theta}^B(\mathbf{x}_0)$ . As benchmarks we consider two proposals by Shen and Zhu (2008), that look for estimators

of  $\theta(\mathbf{x}_0)$  within the class (17) and assume  $g(RSS) = \exp\left(\frac{c}{2}RSS\right)$  with  $c = (n - a)^{-1}$ ,  $a < n$  and thus  $g(RSS) > 1$ . They consider a minimum- $MSE_s$  estimator  $\hat{\theta}^{SMM}(\mathbf{x}_0)$  and a minimum bias estimator  $\hat{\theta}^{SMB}(\mathbf{x}_0)$ .

Shen and Zhu (2008), section 4, prove that  $\hat{\theta}^{SMM}$  compares favorably in terms of  $MSE_s$  to many alternatives popular in the literature and especially so in small samples. In particular it turns out to be more efficient than the naïve back-transform, the restricted maximum-likelihood (REML) estimator, uniformly minimum variance unbiased estimator (UMVU, see Finney, 1941; Bradu and Mundlak, 1970), the bias-corrected REML estimator (El-Shaarawi and Viveros, 1997). Shen and Zhu (2008) note by simulation that  $\hat{\theta}^{SMB}$  behaves closely to the UMVU estimator in most situations but it is much easier to calculate.

We use a synthetic population similar to that considered in Shen and Zhu (2008) but we extend their setting to include other meaningful scenarios. We assume that, in the population  $y_i \sim \text{Logn}(\beta_0 + \beta_1 x_i, \sigma^2)$  where  $(\beta_0, \beta_1) = (1, 1)$  and the values  $x_i$  are fixed constants and distributed uniformly between 0 and 1. We consider three sample sizes  $n = (10, 20, 50)$  and four different variances  $\sigma^2 = (0.1, 0.25, 0.5, 1)$  that correspond to expected  $R^2$  of  $(0.46, 0.25, 0.15, 0.075)$  in the log scale. With respect to Shen and Zhu (2008), we added the case  $\sigma^2 = 0.1$  to cover a situation where the covariate have a good predictive power; higher values of  $R^2$  are not interesting, since all estimators tend to behave very similarly. We consider estimation of  $\theta(\mathbf{x}_0)$ ,  $\mathbf{x}_0 = (1, x_0)$  for  $x_0 \in \{0, 0.1, 0.2, \dots, 1.2\}$ .

The estimators we compare are  $\hat{\theta}^{SMM}(\mathbf{x}_0)$ ,  $\hat{\theta}^{SMB}(\mathbf{x}_0)$  introduced by Shen and Zhu (2008) and  $\hat{\theta}^B(\mathbf{x}_0)$  for three different choices of the hyperparameters. Specifically we consider  $\hat{\theta}^{B1}(\mathbf{x}_0) = E\{\theta(\mathbf{x}_0)|\mathbf{y}, \lambda = 0, \gamma = \gamma_m, \delta = \varepsilon\}$  as suggested in section 4.1;  $\hat{\theta}^{B2}(\mathbf{x}_0) = E\{\theta(\mathbf{x}_0)|\mathbf{y}, \lambda = \lambda_{opt1}, \gamma = \gamma_m, \delta = \varepsilon\}$  and  $\hat{\theta}^{B3}(\mathbf{x}_0) = E\{\theta(\mathbf{x}_0)|\mathbf{y}, \lambda = \lambda_{opt2}, \gamma = \gamma_m, \delta = \varepsilon\}$  according to the  $MSE_s$ -optimizing argument of section 4.2, with non- $\mathbf{x}_0$ -specific and  $\mathbf{x}_0$ -specific  $\lambda$ s, respectively.

The Monte Carlo exercise is based on  $T = 500,000$  replicates. Computations are made by the software R (version 3.0.2).

[Figure 1 about here.]

Results are summarized in Figure 1 and 2. Specifically, in Figure 1 we reported, for the various  $\mathbf{x}_0$ , the ratios of the  $MSE_s$  of the various estimators to the those of the MSE-optimal estimator of (Zellner, 1971, formula 3.9) that assumes  $\sigma^2$  known and is therefore useful as a benchmark but does not represent an alternative to the estimators being considered. We note that for small and moderate sample sizes the choice of hyperparameters has a relevant impact on the frequentist

performances of the various predictors. This is true with the exception of  $\gamma$ : provided it is chosen according to (15) in order to guarantee the existence of posterior means and variances, different choices of its value over a reasonable range does not affect the frequentist properties of Bayes estimators considerably. The choice parameter  $\lambda$ , or equivalently that of  $\delta$  is the one that really impacts on the frequentist properties of the Bayes estimators. Moreover, we note that the posterior distribution of  $\theta(\mathbf{x}_0)$  under the various prior choices, tracks closely that of the associated Bayes estimators with respect to repeated sampling.

The three Bayes estimators and the two proposed by Shen and Zhu (2008) perform differently when  $\sigma^2$  is large and much less so when it is small. In relative terms, the impact of sample size  $n$  looks minor; when  $\sigma^2 = 0.1$  all estimators being considered perform quite closely to that of Zellner (1971) under known  $\sigma^2$  even in small samples, with the only partial exception of  $\hat{\theta}^{B1}(\mathbf{x}_0)$ .

The Bayes estimators based on frequentistically optimal choice of hyperparameters, i.e.  $\hat{\theta}^{B2}(\mathbf{x}_0)$  and  $\hat{\theta}^{B3}(\mathbf{x}_0)$  perform closely to  $\hat{\theta}^{SMM}(\mathbf{x}_0)$  for all  $\mathbf{x}_0$  ranging from 0.1 to 0.9. Values outside this range are associated to  $h_{00} > \frac{1}{3}$  when  $n = 10$  while  $h_{00} < \frac{1}{3}$  in all cases for other sample sizes. For ‘extreme’  $\mathbf{x}_0$ ,  $\hat{\theta}^{B2}(\mathbf{x}_0)$  outperforms all competitors, including  $\hat{\theta}^{SMM}(\mathbf{x}_0)$ .

The Bayes estimator  $\hat{\theta}^{B1}(\mathbf{x}_0)$  is less efficient in terms of  $MSE_s$  especially when  $\sigma^2$  is large and, in relative terms, regardless of the sample size. It performs comparably to others when  $\sigma^2$  is small (i.e.  $R^2$  ‘large’).

In Figure 2 we plotted the ratio of the Bias to square root of the  $MSE_s$  of each estimator. In this case we included the mentioned estimator of Zellner (1971) (notation:  $\hat{\theta}^{ZEL}(\mathbf{x}_0)$ ) in the set of those being compared.

From Figure 2 we note that  $\hat{\theta}^{SMB}(\mathbf{x}_0)$  has a negligible bias in all cases, while all the remaining estimators exhibit a negative bias except  $\hat{\theta}^{B1}(\mathbf{x}_0)$  whose bias is always positive. When estimating  $\theta(\mathbf{x}_0)$ , a smaller  $MSE_s$  than that of the UMVU estimator can be obtained only at the price of a negative bias, that is accepting of systematically underestimating the target parameter.

The bias is relevant for all estimators except  $\hat{\theta}^{SMB}(\mathbf{x}_0)$  and can reach nearly half of the square root of the  $MSE_s$  for ‘extreme’  $\mathbf{x}_0$ .

[Figure 2 about here.]

The estimator  $\hat{\theta}^{B3}(\mathbf{x}_0)$  exhibit a bias that tracks that of  $\hat{\theta}^{ZEL}(\mathbf{x}_0)$  closely when  $h_{00} < \frac{1}{3}$  but with a less variable ‘share’ of the square root of the  $MSE_s$  over the range of the  $\mathbf{x}_0$ . For  $h_{00} > \frac{1}{3}$  its bias is a considerably smaller part of the square root  $MSE_s$  when compared to  $\hat{\theta}^{ZEL}(\mathbf{x}_0)$ ; in the same setting  $\hat{\theta}^{B3}(\mathbf{x}_0)$  is more biased than  $\hat{\theta}^{SMM}(\mathbf{x}_0)$ : its greater efficiency seems to bought at

the price of some additional bias.

As far as  $\hat{\theta}^{B2}(\mathbf{x}_0)$  is concerned, we note that the shape of the ratio of the bias on the root  $MSE_s$  changes with  $\mathbf{x}_0$  differently from the same quantity calculated for  $\hat{\theta}^{B3}(\mathbf{x}_0)$ : heavier for intermediate values, close to 0 on the right and left extremes. Although designed to minimize  $MSE_s$  the choice of the hyperparameters is in this case insensitive to  $\mathbf{x}_0$ . When  $h_{00}$  is large, we end up choosing a  $\lambda$  that is intermediate between 0 (that leads to positive bias) and  $\lambda_{opt2}$  that entails a pronounced negative bias.

We consider also the frequentist coverage of posterior probability intervals based on quantiles (a  $1 - \pi$  interval ( $\pi \in (0, 1)$ ), will range between the quantiles  $\pi/2$  and  $1 - \pi/2$  of the posterior distribution). The small frequentist bias of  $\hat{\theta}^{B1}(\mathbf{x}_0)$  leads to coverage probabilities close to the nominal ones, while performances of  $\hat{\theta}^{B2}(\mathbf{x}_0)$  and  $\hat{\theta}^{B3}(\mathbf{x}_0)$  are less good, at least in some cases. A more detailed account of the simulation results on the frequentist coverage of posterior probability intervals can be found in Appendix B.

## 6 An application to real accelerated life testing data

In this section we present an application of the methodology illustrated in previous sections to a real example from the field of accelerated life testing. We consider a data set discussed in Upadhyay and Peshwani (2008), but initially reported in Nelson (1984). The data consists on a sample of specimens for which a pseudo stress factor  $x$  (the specimen's Young's modulus times its strain) and the number of test cycles  $y$  are measured. The sample points in the original application are 26, but as we are not interested in censoring problems, we consider just 22 uncensored observations. We use 19 out of them as a 'training sample', while three (labelled as  $a$ ,  $b$ ,  $c$  and corresponding to observations 13, 22, 1 in (Upadhyay and Peshwani, 2008, table 2) are left out and used for conditional mean estimation. These observations are chosen as follows: we generated all possible samples that can be obtained excluding 3 out the 22 of the sample and calculate for each of these cross-validation samples the leverage scores for the out-of-sample units. For each of the 22 observations in the original sample, we calculated the median leverage score over all cross-validation samples in which they are not included. We select the three observation with the minimum, average and maximum median leverage score.

The model that we consider is

$$\log y_i = \beta_0 + \beta_1 \log x_i + \beta_2 (\log x_i)^2 + e_i$$

with  $i = 1, \dots, n = 19$  and  $e_i \stackrel{ind}{\sim} N(0, \sigma^2)$ . We assume that this model holds also for the out-of-sample units. Upadhyay and Peshwani (2008) study alternative models, where  $\sigma^2$  is allowed to depend on  $x$ . Here, for simplicity, we do not consider heteroskedastic regression models on the log-scale; this choice is acceptable as the hypothesis of homoskedasticity is not rejected at  $\alpha = 0.05$  significance level using Breusch-Pagan or other popular homoskedasticity tests. The three selected observations for out-of-sample the estimation of  $\theta(\mathbf{x}_a), \theta(\mathbf{x}_b), \theta(\mathbf{x}_c)$  have leverage scores equal to  $h_{aa} = 0.087, h_{bb} = 0.137, h_{cc} = 0.552$ .

We assume  $p(\boldsymbol{\beta}, \sigma^2) = p(\sigma^2)$ ,  $p(\sigma^2) = GIG(\lambda, \delta, \gamma)$  and compare alternative choices of hyperparameters. Posterior distributions are in known form. Required calculations may be performed using for instance the package `ghyp` (Breyman and Lüthi, 2013) running under R.

[Figure 3 about here.]

Posterior densities of  $p\{\theta^b(\mathbf{x}_a)\}, p\{\theta^b(\mathbf{x}_b)\}, p\{\theta^b(\mathbf{x}_c)\}$ ,  $b = B1, B2, B3$  are plotted in figure 3 (left). The vertical line corresponds to the actual value of the dependent variable. The three posterior distributions tend to be close when the leverage is small ( $\mathbf{x}_a$  and  $\mathbf{x}_b$  cases), while in the case  $\mathbf{x}_c$  they are different and  $p\{\theta^{B3}(\mathbf{x}_c)\}$  is much more peaked, consistently with  $\hat{\theta}^{B3}(\mathbf{x}_c)$  being a more  $MSE_s$ -efficient estimator; we note also that  $p\{\theta^{B1}(\mathbf{x}_*)\}$  is in all cases shifted to the right and has a heavier right tail, as it may be expected from  $\hat{\theta}^{B1}(\mathbf{x}_*)$  being a positively biased estimator, in contrast with the other two Bayes estimators that are negatively biased.

To shed more light on the impact hyperparameters of  $p(\sigma^2)$  may have on posterior inference when the sample size is small, in the right panel of figure 3 we plotted the posterior densities of the regression parameter  $\beta_1$  under the alternative choices considered so far for the prior on  $\sigma^2$  and that we label for short as  $B1, B2, B3(\mathbf{x}_a), B3(\mathbf{x}_b), B3(\mathbf{x}_c)$ , along with that obtained under the reference prior  $p(\sigma^2) \propto \sigma^{-2}$ . As expected from (14), we have that when  $\lambda = 0, \delta = \varepsilon$  ( $B1$ ), the posterior density is very close to that obtained under the reference prior. It appears slightly less peaked; this is also expected as the prior choice  $B1$  is characterized by lighter tails than the reference prior as the first implies finite posterior variance for the conditional mean while the second does not. Posterior densities associated to  $MSE_s$ -optimal choices are more peaked (although characterized by the same location); this effect is more pronounced as far as the  $h_{00}$  value substituted in (21) gets high.

In summary, using the theory from section 4, we learned that a careful choice of hyperprior parameters may lead to very efficient estimators of  $\theta(\mathbf{x}_0)$  in terms of  $MSE_s$ ; when the sample size is small the pursuit of this goal may on the other hand have undesired impacts on other aspects of



inference, such as a deterioration of the frequentist properties of posterior distribution summaries.

## 7 Concluding remarks

In this paper we considered the problem of estimating the conditional mean of the dependent variable in the log-normal linear regression model. We showed that many prior distributions for the variance parameter (on the log-scale) lead to posteriors without finite moments, thus impossible to summarize using common loss functions. We motivated the recourse to the generalized inverse gaussian prior, derived the constraints on its parameter that allow for finite posterior moments of a pre-specified order and discussed how to choose the remaining parameters of this distribution. We introduced ‘reference’ choices that lead to posterior distributions for the regression coefficients very close to those obtained under the usual reference prior; nonetheless the posterior expectation of the conditional mean, when considered as frequentist ‘Bayes’ estimator, is suboptimal with respect to known alternative estimators. Moreover, we adopted a frequentist optimality criterion, i.e. minimizing the  $MSE_s$  of the (approximated) posterior mean. The search for estimators of the conditional mean with minimum  $MSE_s$  has a long tradition in the literature, from Finney (1941) to Shen and Zhu (2008). In this line we obtained a ‘Bayes’ estimator that compares favourably to known alternatives in the frequentist literature.

The model we considered is very simple and can be generalized in many directions, namely assuming dependent and/or heteroskedastic residuals. Nonetheless, the problem we focused on, the non-existence of posterior moments for the conditional mean under many popular priors for the residuals’ variance (on the log-scale), affects more general models in the same way as it affects ours, i.e. every time we need to take the *exp* transformation of a  $t$  distribution in the estimation process. We developed a solution in an analytically tractable case, leaving extensions to more general models for future research.

## Supporting information

Additional Supporting Information may be found in the online version of this article.

**Appendix A** contains the proofs of theorems 3.1, 3.2, 4.1, 4.2, 4.3.

**Appendix B** is about the frequentist coverage of posterior probability intervals, discussed by means of a simulation exercise as mentioned in section 5.

## Acknowledgments

The research of Enrico Fabrizi was partially supported by the PRIN 2012 project ‘Household wealth and youth unemployment: new survey methods to meet current challenges’.

## References

- Ahn, J.O. & Ku, J.H. (2006) The relationship between Serum Prostate-Specific Antigen Levels and Body Mass Index in Healthy Younger Men, *Urology*, **68**, 570–574.
- Barndorff-Nielsen, O.E. (1977), Exponentially decreasing distributions for the logarithm of particle size, *Proceedings of the Royal Statistical Society*, series A, **353**, 401–419.
- Bibby, B.M. & Sørensen, M. (2003), Generalized hyperbolic and inverse Gaussian distributions: limiting cases and approximation of processes, in *Handbook of Heavy Tailed Distributions in Finance* (ed. S.T. Rachev), 211–248. Elsevier Science B.V., Amsterdam.
- Bradu, D. & Mundlak, Y. (1970) Estimation in lognormal linear models. *J. Amer. Statist. Assoc.* **65**, 198–211.
- Breymann, W. & Lüthi, D. (2013), `ghyp`: a package on Generalized Hyperbolic distributions, available at [cran.r-project.org/web/packages/ghyp/index.html](http://cran.r-project.org/web/packages/ghyp/index.html).
- Eberlein, E. & von Hammerstein E.A. (2004), Hyperbolic processes in finance, in *Seminar on Stochastic Analysis, Random Fields and Applications IV, Progress in Probability*, (eds. R.C. Dalang, M. Dozzi, F. Russo), 221–264. Birkhäuser Verlag, Berlin.
- El-Shaarawi, A.H. & Viveros, R. (1997) Inference about the mean in log-regression with environmental applications, *Environmetrics*, **8**, 569–582.
- Fabrizi, E. & Trivisano, C. (2012) Bayesian estimation of log-normal means with finite quadratic expected loss, *Bayesian Anal.*, **7**, 975–996.
- Finney, D.J. (1941) On the distribution of a variate whose logarithm is normally distributed, *J. R. Stat. Soc.*, **7** Suppl., 144–161.
- Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models, *Bayesian Anal.*, **1**, 515–533.

- Griffin, J.E. & Steel, M.F.J. (2010) Bayesian nonparametric modelling with the Dirichlet process regression smoother, *Statist. Sinica* **20**: 1507–1527.
- Karlberg, F. (2000) Population total prediction under a lognormal superpopulation model, *Metron*, **LVIII**, 53–80.
- Li, Q. & Lin, N. (2010) The Bayesian elastic net, *Bayesian Anal.*, **5**, 151–170.
- Mankiw, N.G., Romer, D. & Weil, D. (1992) A contribution to the empirics of economic growth, *Q. J. Econ.*, **107**, 407–437.
- Nelson, W. (1984) Fitting of fatigue curves with non-constant standard deviation to data with run outs, *J. Test. Eval.*, **12**, 69–77.
- Nelson, W. (1990) *Accelerated testing: statistical models, test plans and data analysis*, Wiley, New York.
- Olin, A.C., Bake, B. & Toren, K. (2007) Fraction of Exhaled Nitric Oxide at 50 mL/ $\mu$ . Reference Values for Adult Lifelong Never-Smokers, *Chest*, **131**, 1852–1856.
- Rukhin, A.L. (1986) Improved estimation in lognormal models, *J. Amer. Statist. Assoc.*, **81**, 1046–1049.
- Shen H. & Zhu, Z. (2008) Efficient mean estimation in log-normal linear models, *J. Statist. Plann. Inference*, **138**, 552–567.
- Tabane, L. & Haq, SM. (1999) Prediction from a normal model using a generalized inverse Gaussian prior, *Statist. Papers*, **40**, 175–184.
- Upadhyay, S.K. & Peshwani M. (2008) Posterior analysis of lognormal regression models using the Gibbs sampler, *Statist. Papers*, **49**, 59–85.
- Zellner, A., Kmenta, J. & Drèze, J. (1966) Specification and estimation of Cobb-Douglas production function models, *Econometrica*, **34**, 784–795.
- Zellner, A. (1971) Bayesian and non-Bayesian analysis of the log-normal distribution and log-normal regression, *J. Amer. Statist. Assoc.*, **66**, 327–330.

## List of Figures

1	Ratios of the $MSE_s$ of various estimators of $\theta(\mathbf{x}_0)$ to that of minimum- $MSE_s$ estimator suggested by Zellner (1971) that assumes $\sigma^2$ known . . . . .	19
2	Ratios of the Bias of various estimators of $\theta(\mathbf{x}_0)$ to the square root of their $MSE_s$ . . . . .	20
3	Posterior densities of conditional means $\theta(\mathbf{x}_a)$ , $\theta(\mathbf{x}_b)$ , $\theta(\mathbf{x}_c)$ (left); Posterior densities of $\beta_1$ under different priors for the $\sigma^2$ parameter (right). . . . .	21

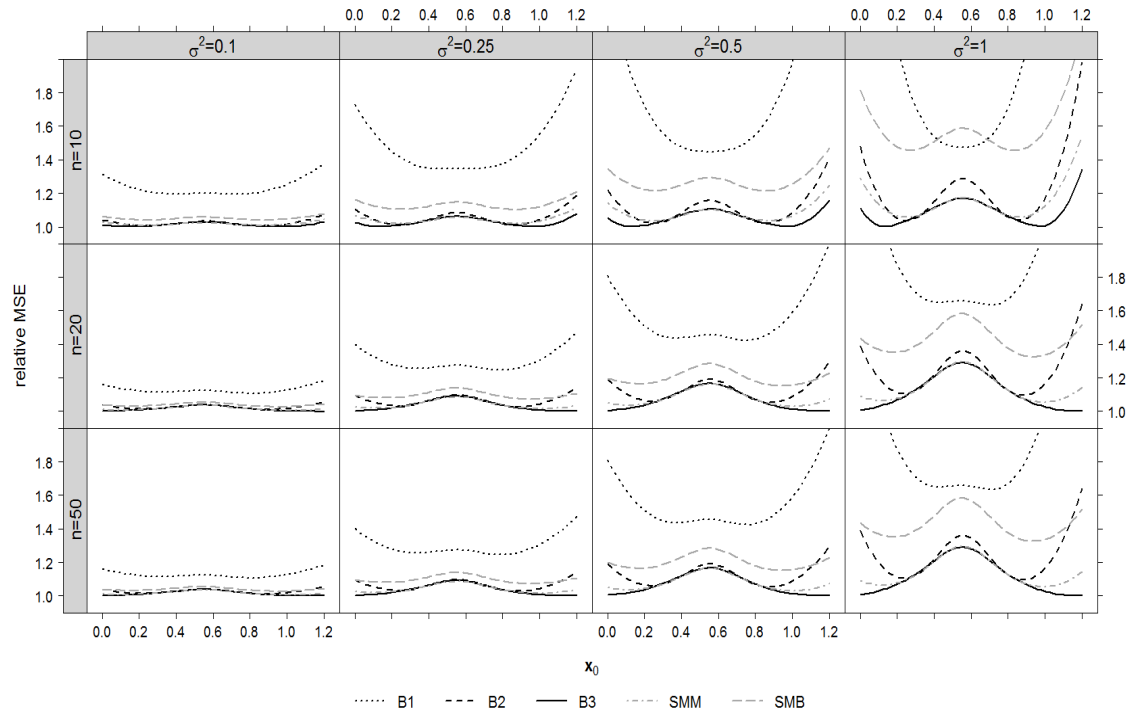


Figure 1: Ratios of the  $MSE_s$  of various estimators of  $\theta(\mathbf{x}_0)$  to that of minimum- $MSE_s$  estimator suggested by Zellner (1971) that assumes  $\sigma^2$  known

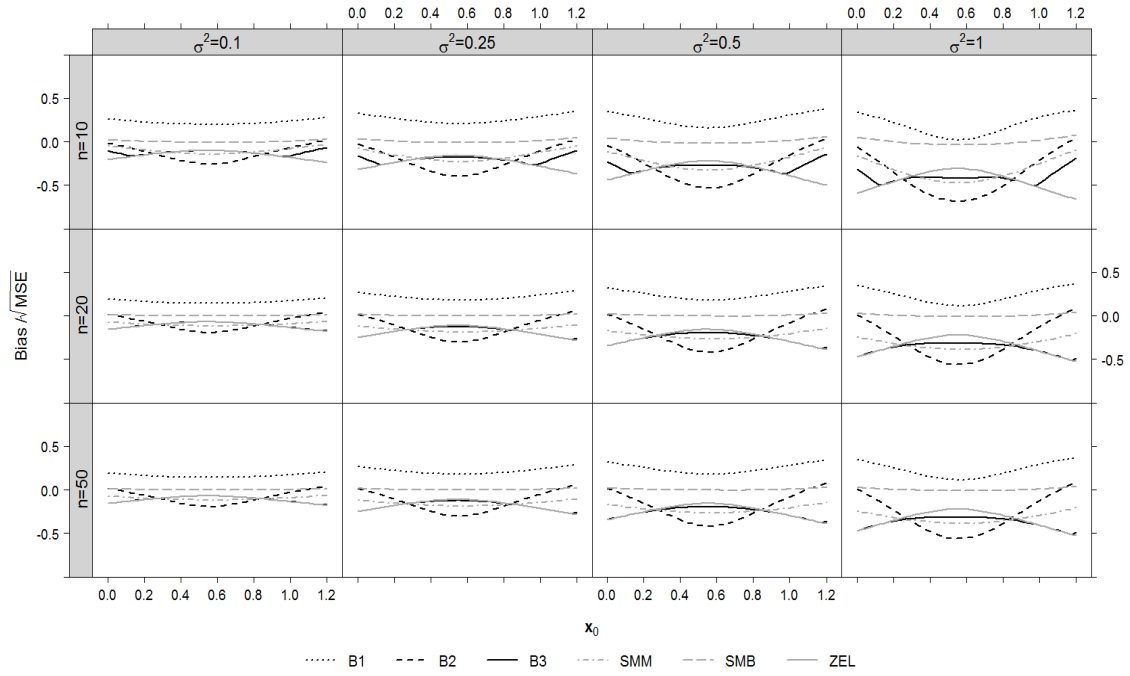


Figure 2: Ratios of the Bias of various estimators of  $\theta(\mathbf{x}_0)$  to the square root of their  $MSE_s$

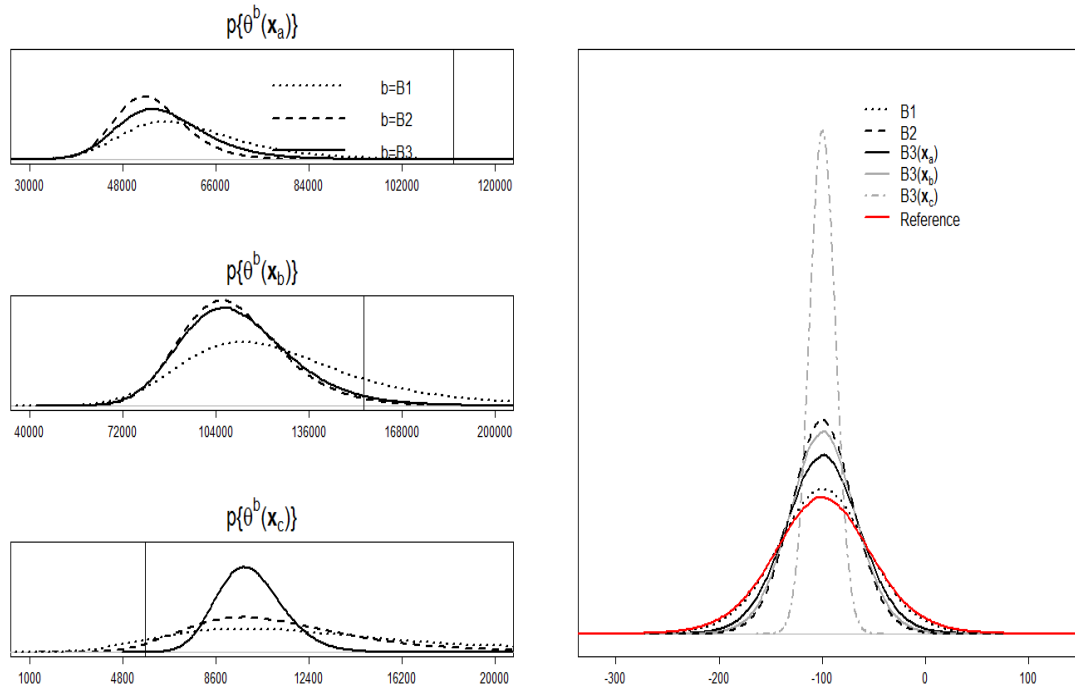


Figure 3: Posterior densities of conditional means  $\theta(\mathbf{x}_a)$ ,  $\theta(\mathbf{x}_b)$ ,  $\theta(\mathbf{x}_c)$  (left); Posterior densities of  $\beta_1$  under different priors for the  $\sigma^2$  parameter (right).