# Tricks with Metrics: Combining Statistics for Improved Inference in Regression Analysis

Pierre Nguimkeu*
Georgia State University

September 2023

**Abstract**

In maximum likelihood methods, the three classical tests statistics are often unreliable for inference in small samples even under correct model specification. In this paper, I discuss how the likelihood-ratio and Wald tests statistics can be combined to obtain highly improved parameter inference in regression models with small samples. I consider modifications obtained from both the Barndorff-Nielsen and the Lugannani and Rice likelihood approximations, and I show how they can produce highly accurate parameter inference in a general (possibly nonlinear) regression model with possibly non-spherical disturbances. I discuss the underlying theory and provide Monte Carlo simulations demonstrating the superior accuracy of the proposed procedures over the first-order classical likelihood methods (i.e., the signed log-likelihood ratio test and the Wald test). An empirical application to a regression model of mobile money ("M-pesa") adoption in Kenya is provided as an illustration of the usefulness of these methods in practice.

**Keywords:** Small sample inference; Likelihood analysis, Third-order approximation.

**JEL Codes:** C12; C15; C18.

## 1 Introduction

Regression analysis is a central technique for empirical work in social sciences in general and in Economics in particular. Standard asymptotic tests statistics such as the likelihood-ratio test statistic, the Wald test statistic or the Score test statistic are often used to perform inference about parameters of interest, under some distributional assumption. These likelihood based tests are appealing because they offer a general and straightforward method for obtaining p-values and confidence intervals. However, in some settings (e.g., nonlinear multivariate regressions and/or regressions with nonnormal errors), both the Score, the Wald and the Likelihood-ratio tests can have poor properties when the sample size is small, or in general, when the average information available per parameter is limited. Many examples are available to illustrate the failure of these methods - usually referred to as first-order approximations - in models with small samples and/or large numbers of nuisance parameters. Reviews of such examples can be found in Barndorff-Nielsen and

*Department of Economics, Georgia State University, 55 Park Place NE, Suite 600, Atlanta, GA 30303, USA; Email: nnguimkeu@gsu.edu.

Cox (1994), Elkantassi et al. (2023), Fraser and Reid (1995), Severini (2000), Butler (2007), Brazzale and Davison (2008), Brazzale, Davison and Reid (2007), Fraser (2017).

For more accurate inference, refinements are needed to improve upon the first-order approximations, find the distribution of the relevant statistic in the presence of small samples, and/or properly account for nuisance parameters in the multi-parameter setting. This paper describes simple adjustments (tricks) that combine the Log-likelihood ratio statistic and the Wald statistic (metrics) to obtain modified test statistics that are much more accurate in small samples than these classical ones. These adjustments are based on results in higher-order asymptotics developed by Lugannani and Rice (1980), Barndorff-Nielsen (1983, 1986), Fraser (1990), Fraser, Reid and Wu (1999) and Skovgaard (1996). The development of these analytical approximations have led to a theory of near-exact inference based on small samples from parametric models, and they not only provide modifications to well-established approaches which result in more accurate inferences, but also give insight on when to rely upon classical first-order methods. Their theoretical basis is the saddlepoint and related approximations that emerged in Daniels (1954, 1987) and further developments have been well described by Reid (1988, 1995, 2003). These methods are highly accurate in many situations especially when dealing with small samples or large numbers of nuisance parameters, but have nevertheless been under-used in Econometrics, Finance and other Social Sciences beyond the field of Computational Statistics, compared to simulation procedures such as the Bootstrap. The reason may have been the technical details of the methods or the lack of suitable software and computer power. But recent computational advances have overcome these issues, and can allow likelihood-based small-sample parametric asymptotics to be widespread.

In this paper, I focus on the use of higher-order parameter inference methods in a general regression model with possibly nonlinear location and possibly nonnormal and/or nonspherical errors. This includes several special classes of models, e.g., binary choice models such as the logit model, nonnormal linear regression models and nonlinear regression models with heteroscedastic normal errors, seemingly unrelated regression models and others.[1] I also provide computer programs for the implementation of these methods. The objective is to make higher-order inference for such models better known and available for use by applied researchers who do not necessarily have a command of the technical details. Using real data from mobile money adoption in Kenya the proposed methods are applied to a logistic growth model of technology diffusion at its early adoption stage where only few data points are available. Some background on the regression model is given in Section 2, and a review of the relevant likelihood asymptotic theory is given in Section 3. The likelihood-based results for the general regression model are developed in Section 4. Numerical examples involving Monte Carlo simulations and real data applications are provided in Sections 5. Concluding remarks are given in Section 6.

[1] The seemingly unrelated regression model is discussed by Fraser, Rekkas and Wong (2005).

## 2 Regression Models and Classical Tests

Consider a general regression model given by

$$\mathbf{y} = \mathbf{g}(\mathbf{x}, \boldsymbol{\beta}) + \Omega\boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is an $n$-vector of the dependent variable, $\mathbf{x}$ is an $n \times K$ matrix of regressors, $\boldsymbol{\epsilon}$ is an $n$-vector of possibly nonnormal errors with a distribution $f(\boldsymbol{\epsilon})$, and $\mathbf{g}(\mathbf{x}, \boldsymbol{\beta})$ is an $n$-vector of possibly nonlinear smooth functions of $\mathbf{x}$ and a $K$-parameter vector $\boldsymbol{\beta}$, and $\Omega$ is a scaling matrix which could potentially depend on data $\mathbf{x}$ and/or parameters $\boldsymbol{\gamma}$ to allow for possibly heteroskedasticity and/or serial correlation.[2] In this model, we are often interested in testing one of the scalar parameters $\beta_1, \ldots, \beta_K$ denoted $\beta_k$ (or one of the scalar parameters in the matrix $\Omega$). We could also be interested in jointly testing several parameters of this model. The latter can then be easily obtained by testing them sequentially, holding the preceding ones fixed. In the classical approach, inference for the parameter $\beta_k$ is provided in the form of a $p$-value, $p(\beta_k)$, that gives the probability position of the data relative to the value $\beta_k$ for the parameter. A test of the parameter value $\beta_k$ consists in seeing whether the p-value $p(\beta_k)$ is close to 0 or close to 1, and a $100(1-\alpha)\%$ confidence interval is obtained by inverting the $p$-value function as follows

$$[\hat{\beta}_k^L, \quad \hat{\beta}_k^U] = \left[\min\left\{p^{-1}(1-\alpha/2), p^{-1}(\alpha/2)\right\}, \quad \max\left\{p^{-1}(1-\alpha/2), p^{-1}(\alpha/2)\right\}\right]. \tag{2}$$

When the model is linear with normal homoskedastic errors, ie $\mathbf{g}(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{X}'\boldsymbol{\beta}$, where $\mathbf{X} = [\mathbf{i} \;\; \mathbf{x}]$ with $\mathbf{i} = [1 \ldots 1]'$, $\Omega = \sigma \mathbf{I}_n$ and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_n)$, then

$$p(\beta_k) = F_{n-K}(t_k) \tag{3}$$

and

$$t_k = \frac{(\hat{\beta}_k - \beta_k)}{se(\hat{\beta}_k)} \tag{4}$$

where $\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\mathbf{y}$, $s^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/(n-K)$, $se(\hat{\beta}_k)$ is the squared-root of the $(k,k)$ diagonal element of $s^2(\mathbf{X}'\mathbf{X})^{-1}$, and $F_{n-K}$ is the Student distribution function with $n-K$ degrees of freedom.

In the general case, the regression model (1) does not typically have an exact test statistics for inference on $\beta_k$ or an exact distribution to produce the $p$-value, $p(\beta_k)$, as given by Equation (3). A common approach is to rely on large sample theory results such as the central limit theorem and use likelihood-based classical first-order test statistics such as the signed likelihood ratio departure $r$, the Wald departure $q$ or the score statistic $z$, given by:

$$r = \operatorname{sgn}(\hat{\beta}_k - \beta_k)[2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_k)\}]^{1/2} \tag{5}$$

$$q = (\hat{\beta}_k - \beta_k)(\hat{j}^{kk})^{-1/2} \tag{6}$$

[2]This set up is similar to the one discussed in Nguimkeu and Rekkas (2011). However, the present framework is more general since the error density $f(\boldsymbol{\epsilon})$ is arbitrary, and inference concerns any scalar parameter in the model including both the mean function and the variance parameters.

$$z = l_k(\hat{\boldsymbol{\theta}}_k)(\hat{j}^{kk})^{1/2} \tag{7}$$

where $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_K, \boldsymbol{\gamma}')'$ is the full parameter vector, $l(\boldsymbol{\theta}) = \ln f(\mathbf{y}, \boldsymbol{\theta})$ is the log-likelihood function, $l_k(\boldsymbol{\theta}) = \partial \ln f(\mathbf{y}, \boldsymbol{\theta})/\partial\beta_k$ is the score for $\beta_k$, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_K, \boldsymbol{\gamma})$, $\hat{\boldsymbol{\theta}}_k$ is the constrained maximum likelihood estimator when the component $\beta_k$ is held at its hypothesized value, $\hat{j}^{kk}$ is the $(k,k)$ element of the inverse $\hat{j}^{-1}$ of the observed information matrix $\hat{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'} = \partial^2 l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. For implementation, reliable optimizing routines are needed for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_k$, especially when the error density $f(\boldsymbol{\epsilon})$ has long tails. The statistics $r$, $q$ and $z$ have standard normal distributions up to the order $O(n^{-1/2})$. The $p$-values for inference on $\beta_k$ are therefore given by $\Phi(r)$, $\Phi(q)$ and $\Phi(z)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal.[3] However, when the sample size is small these first-order approximations are often inaccurate, especially in complex models.

# 3 Modified Likelihood-Based Tests

The testing procedures discussed in this paper for the general regression model (1) are modifications of the directed log-likelihood statistic (i.e. the signed square-root of the log-likelihood ratio statistic given in (5)). The keys to refining the approximate behavior of this likelihood quantity are two higher-order density approximations: Barndorff-Nielsen's (1983, 1986) formula and the tangent exponential model developed by Fraser, Reid and Wu (1999). The former gives the density of the maximum likelihood estimate at the observed data and at other points having the same value of an ancillary statistic. The latter is an exponential model whose distribution function at the observed data differs from that of the conditional model by $O(n^{-3/2})$ under the observed conditioning (see, e.g., Fraser, Andrews & Wong, 2005).

Denote by $\psi$ the scalar parameter of primary interest and $\boldsymbol{\lambda}$ a vector of parameters that are not of direct concern (nuisance parameters), i.e. $\boldsymbol{\theta} = (\psi, \boldsymbol{\lambda}')'$. This partitioning entails corresponding splits of the score vector $l_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ into $l_\psi(\boldsymbol{\theta})$ and $l_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$, and of the observed information matrix $\hat{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\boldsymbol{\theta})$ into the sub-matrices $\hat{j}_{\psi\psi}(\boldsymbol{\theta})$, $\hat{j}_{\psi\boldsymbol{\lambda}'}(\boldsymbol{\theta})$, $\hat{j}_{\boldsymbol{\lambda}\psi}(\boldsymbol{\theta})$ and $\hat{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta})$. Third-order likelihood theory produces a metric of the departure of a data point from a value of $\psi$ and also produces the corresponding $p$-value. The departure measure does not arise explicitly, but is analogous to the $t$ statistic in (4). It is given by the modified signed log-likelihood statistic defined by

$$r^*(\psi) = r(\psi) - r(\psi)^{-1} \ln\left(\frac{r(\psi)}{Q(\psi)}\right) \tag{8}$$

where $r(\psi)$ is the signed log-likelihood ratio statistic

$$r(\psi) = \operatorname{sgn}(\hat{\psi} - \psi)\left[2\left\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_\psi)\right\}\right]^{1/2} \tag{9}$$

and $Q(\psi)$ is a special Wald-type departure metric defined by

[3] In practice, the approximate normality of $r$ is usually better than that of $q$ and $z$.

$$Q(\psi) = \text{sgn}(\hat{\psi} - \psi)|\chi(\hat{\boldsymbol{\theta}}) - \chi(\boldsymbol{\theta})| \left\{ \frac{|\hat{j}_{\boldsymbol{\varphi}\boldsymbol{\varphi}'}|}{|\hat{j}_{(\boldsymbol{\lambda}\boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}_\psi)|} \right\}^{1/2} \tag{10}$$

where $\hat{j}_{\boldsymbol{\varphi}\boldsymbol{\varphi}'}$ and $\hat{j}_{(\boldsymbol{\lambda}\boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}_\psi)$ are the observed information matrix and observed nuisance information matrix, respectively, calculated in the nominal parametrization scale. More specifically, they can be obtained as follows:

$$\hat{j}_{\boldsymbol{\varphi}\boldsymbol{\varphi}'}(\hat{\boldsymbol{\theta}}) = |\hat{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}})||\boldsymbol{\varphi}_{\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}})|^{-2} \tag{11}$$

$$\hat{j}_{(\boldsymbol{\lambda}\boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}_\psi) = |\hat{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta}_\psi)||\boldsymbol{\varphi'}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_\psi)\boldsymbol{\varphi}_{\boldsymbol{\lambda}'}(\hat{\boldsymbol{\theta}}_\psi)|^{-1} \tag{12}$$

In the above formula, $\chi(\boldsymbol{\theta})$ acts as a scalar canonical parameter for a one parameter marginal model for testing the interest parameter $\psi$

$$\chi(\boldsymbol{\theta}) = \frac{\psi_{\boldsymbol{\varphi}'}(\boldsymbol{\theta})}{|\psi_{\boldsymbol{\varphi}'}(\boldsymbol{\theta})|}\boldsymbol{\varphi}(\theta), \tag{13}$$

where $\psi_{\boldsymbol{\varphi}'}(\boldsymbol{\theta}) = \dfrac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\varphi}'} = \left(\dfrac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\left(\dfrac{\partial\boldsymbol{\varphi}'(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)^{-1}$. Now, all that is needed is the nominal reparameterization $\boldsymbol{\varphi}(\boldsymbol{\theta})$ specific to the data point $\mathbf{y}$, given by

$$\boldsymbol{\varphi}'(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta};\mathbf{y})}{\partial\mathbf{V}} = \frac{\partial l(\boldsymbol{\theta};\mathbf{y})}{\partial\mathbf{y}'}\mathbf{V}. \tag{14}$$

The vectors in $\mathbf{V}$ indicate how $\mathbf{y}$ responds to changes in $\boldsymbol{\theta}$ and can be constructed coordinate by coordinate from a full $n$-dimensional pivotal quantity $\mathbf{q}(\boldsymbol{\theta};\mathbf{y})$ that has $\boldsymbol{\theta}$-free distribution as follows:

$$\mathbf{V} = \left.\frac{\partial\mathbf{y}}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \left\{\frac{\partial\mathbf{q}(\boldsymbol{\theta};\mathbf{y})}{\partial\mathbf{y}'}\right\}^{-1}\left\{\frac{\partial\mathbf{q}(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\theta}'}\right\}. \tag{15}$$

If the model is exponential then $\boldsymbol{\varphi}(\boldsymbol{\theta})$ can be taken to be any version of the canonical parameter. In general, $\boldsymbol{\varphi}(\boldsymbol{\theta})$ is a gradient of the log-likelihood taken in directions that conform to an approximate ancillary that describes the model structure locally.

The highly accurate likelihood-based p-value $p(\psi)$ employed to test the value of a scalar parameter of interest $\psi(\theta) = \psi$ can then be obtained using either of the asymptotically equivalent expressions:

$$\Phi(r) + \phi(r)\left\{\frac{1}{r} - \frac{1}{Q}\right\} \tag{16}$$

and

$$\Phi(r^*) = \Phi\left(r - r^{-1}\ln\left\{\frac{r}{Q}\right\}\right), \tag{17}$$

where $\phi(\cdot)$ is the probability density function (pdf) of the standard normal. Confidence intervals at $100(1-\alpha)\%$ are then obtained by inverting these p-value functions are given in Equation (2). Expression (16) is known as the Lugannani and Rice (1980) approximation and expression (17) is known as the Barndorff-Nielsen (1991) approximation based on the modified signed log-likelihood statistic $r^*$. As shown by these authors, these two approximations, which are combinations of the classical statistics $r$ and $Q$, converge at a $O(n^{-3/2})$ rate and are thus referred to as third-order approximations. They are superior to the latter two which have the usual $O(n^{-1/2})$ rate of convergence and are hence commonly referred to as first-order approximations.

The above discussion applies in general to continuous response models. For discrete responses, analogous results can be found in Davison, Fraser, and Reid (2006). However, for distributions whose support has a lattice structure or can easily be transformed to a lattice structure (e.g. binomial and Poisson ), the use of a slightly modified form of (16) provides approximations to tail probabilities with error $O(n^{-1})$ ( Severini, 2000; Davison, 2003). Likewise, when the parameter of interest is a vector, Skovgaard (2001) suggests adjusting the likelihood ratio statistic in a way that generalizes $r^*$ to the multiparameter case.

# 4 Improved Inference in Regression Models

We now use the modified likelihood-based methods described above to develop inference procedures for the regression model (1). We assume that the regression function $\mathbf{g}(\mathbf{x}, \boldsymbol{\beta})$ is continuously differentiable and that the error density function $f(\boldsymbol{\epsilon})$ is also continuously differentiable. We assume that the scaling matrix $\Omega$ is positive definite and is indexed by a fixed low-dimensional vector of parameters $\boldsymbol{\gamma}$, that is, $\Omega = \Omega(\boldsymbol{\gamma})$. In practice, this would be the case when $\Omega$ captures, for example, possible heteroskedasticity and/or serial correlation in the model with a well defined covariance structure. Emphasizing this parameterization, the model is rewritten as

$$\mathbf{y} = \mathbf{g}(\mathbf{x}, \boldsymbol{\beta}) + \Omega(\boldsymbol{\gamma})\boldsymbol{\epsilon}. \tag{18}$$

If the parameter of interest on which we wish to perform inference is taken to be $\beta_k$, the full model parameter vector can be rewritten as

$$\boldsymbol{\theta} = (\boldsymbol{\beta'}, \boldsymbol{\gamma'})' = (\beta_k, \boldsymbol{\lambda}')',$$

with $\boldsymbol{\lambda} = (\beta_1, \ldots, \beta_{k-1}, \beta_{k+1} \ldots, \beta_K, \boldsymbol{\gamma}')'$. If the interest is on $\gamma_s$, then $\boldsymbol{\theta} = (\gamma_s, \boldsymbol{\lambda}')'$, and the nuisance parameter is $\boldsymbol{\lambda} = (\boldsymbol{\beta}', \gamma_1, \ldots, \gamma_{s-1}, \gamma_{s+1} \ldots, \gamma_S, )'$. In the discussion below we focus on the former case, but the latter case follows similarly by changing the roles of the scalar parameter of interest. In the numerical results, we illustrate both cases. This framework allows wide generality for the error distribution, including the case of dependent errors (captured by the covariance matrix $\Omega$). We therefore assume that the components $\epsilon_i$ of the vector $\boldsymbol{\epsilon}$ are independent with individual density $f_i(\epsilon) = \exp\{l_i(\epsilon)\}$ and assume that these densities have been centered so that the individual scores $\mathcal{S}_i(\epsilon) = \dfrac{dl_i(\epsilon)}{d\epsilon}$ are 0 at the origin, that is, $\mathcal{S}_i(0) = \left.\dfrac{dl_i(\epsilon)}{d\epsilon}\right|_{\epsilon=0} = 0$. Denote by $\mathbf{e}_i$ the $n$-dimensional vector whose $i^{th}$ element is 1 and all the other elements are 0.

The log-likeliood function of the model has the form

$$l(\boldsymbol{\theta}) = -\log|\Omega| + \sum_{i=1}^{n} l_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y} - \mathbf{g}(\mathbf{x}, \boldsymbol{\beta})]\right) \tag{19}$$

In the normal error model, i.e. $l_i(\epsilon) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\epsilon^2$, the log-likelihood function (19) takes the form

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{1}{2}\log(2\pi) - \log|\Omega| - \frac{1}{2}\sum_{i=1}^{n}[\mathbf{y} - \mathbf{g}(\mathbf{x}, \boldsymbol{\beta})]'\Omega'^{-1}\mathbf{e}_i\mathbf{e}_i'\Omega^{-1}[\mathbf{y} - \mathbf{g}(\mathbf{x}, \boldsymbol{\beta})] \\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}[\mathbf{y} - \mathbf{g}(\mathbf{x}, \boldsymbol{\beta})]'\Sigma^{-1}[\mathbf{y} - \mathbf{g}(\mathbf{x}, \boldsymbol{\beta})], \end{aligned}$$

where the last display follows by noting that $\sum_{i=1}^{n}\mathbf{e}_i\mathbf{e}_i' = \mathbf{I}_n$ and denoting $\Omega\Omega'$ by $\Sigma$. The case of homoscedastic normal errors can then be treated by setting $\Sigma = \Sigma(\sigma) = \sigma^2\mathbf{I}_n$. The heteroskedastic case can be treated by denoting $\boldsymbol{\gamma} = (\sigma, \boldsymbol{\delta}')'$ and setting $\Sigma = \Sigma(\sigma, \boldsymbol{\delta}) = \sigma^2 W(\mathbf{x}, \boldsymbol{\delta})$, where the matrix function $W(\mathbf{x}, \boldsymbol{\delta})$ may depend on the data. The matrix $\Sigma$ can also be defined to handle the case of serially correlated disturbances. For example, for a regression model with first-order autoregressive disturbances (AR(1)) with autocorrelation coefficient $\rho$ and error variance $\sigma^2$, the variance parameter vector is $\boldsymbol{\gamma} = (\sigma, \rho)'$ and an expression for the matrix $\Sigma$, which can be found in Nguimkeu and Rekkas (2011), is given by

$$\Sigma = \Sigma(\sigma, \rho) = \sigma^2 \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1+\rho^2 & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}. \tag{20}$$

Likewise, for a regression model with first-order moving average disturbances (MA(1)) with correlation coefficient $\rho$ and error variance $\sigma^2$, an expression for the matrix $\Sigma$, which can be found in Nguimkeu (2014), is given by

$$\Sigma = \Sigma(\sigma, \rho) = \sigma^2 \begin{pmatrix} 1+\rho^2 & \rho & 0 & \cdots & 0 \\ \rho & 1+\rho^2 & \rho & \cdots & 0 \\ 0 & \rho & 1+\rho^2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho \\ 0 & 0 & \cdots & \rho & 1+\rho^2 \end{pmatrix}. \tag{21}$$

Maximization of the general log-likelihood function (19) over the parameter space produces the maximum likelihood estimator $\hat{\boldsymbol{\theta}} = (\hat{\beta}_k, \hat{\boldsymbol{\lambda}}')'$. This is usually calculated using some iterative procedure applied to the normal equations given by

$$l_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \nabla_{\boldsymbol{\beta}}\mathbf{g}(\mathbf{x},\boldsymbol{\beta})\Omega'^{-1}\mathbf{e}_i\mathcal{S}_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\right) = 0 \tag{22}$$

$$l_{\gamma_s}(\boldsymbol{\theta}) = -\mathrm{tr}\left(\Omega^{-1}\frac{\partial\Omega}{\partial\gamma_s}\right) + \sum_{i=1}^{n}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]'\frac{\partial\Omega'^{-1}}{\partial\gamma_s}\mathbf{e}_i\mathcal{S}_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\right) = 0, \quad s = 1,\ldots,m \tag{23}$$

where $\nabla_{\boldsymbol{\beta}}\mathbf{g}(\mathbf{x},\boldsymbol{\beta}) = \left(\frac{\partial\mathbf{g}(\mathbf{x},\boldsymbol{\beta})'}{\partial\beta_1},\ldots,\frac{\partial\mathbf{g}(\mathbf{x},\boldsymbol{\beta})'}{\partial\beta_k}\right)'$ is the $k\times n$ matrix of partial derivatives of $\mathbf{g}(\mathbf{x},\boldsymbol{\beta})$ and $\frac{\partial\Omega^{-1}}{\partial\gamma_s} = -\Omega^{-1}\frac{\partial\Omega}{\partial\gamma_s}\Omega^{-1}$. The solution to these equations gives the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. The constrained maximum likelihood estimator $\hat{\boldsymbol{\theta}}_k = (\beta_k, \hat{\boldsymbol{\lambda}}')'$ is then computed in the same way but with $\beta_k$ fixed at its hypothesized value and the associated equation $l_{\beta_k}(\boldsymbol{\theta}) = 0$ omitted in the system of equations (22). Likewise, if the interest is in $\gamma_s$, then the constrained maximum likelihood estimator $\hat{\boldsymbol{\theta}}_s = (\gamma_s, \hat{\boldsymbol{\lambda}}')'$ is obtained by jointly solving these equations but with $\gamma_s$ fixed at its hypothesized value and the equation $l_{\gamma_s}(\boldsymbol{\theta}) = 0$ omitted in the system (23).

To compute the information matrix, $\mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}$, we need the second order derivatives given by

$$\begin{aligned} l_{\boldsymbol{\beta}\boldsymbol{\beta}'}(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\nabla_{\boldsymbol{\beta}}\mathbf{g}(\mathbf{x},\boldsymbol{\beta})\Omega'^{-1}\mathbf{e}_i\mathbf{e}_i'\Omega^{-1}\nabla_{\boldsymbol{\beta}'}\mathbf{g}(\mathbf{x},\boldsymbol{\beta})\mathcal{H}_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\right) \\ &\quad - \sum_{i=1}^{n}\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}'}\mathbf{g}(\mathbf{x},\boldsymbol{\beta})\Omega'^{-1}\mathbf{e}_i\mathcal{S}_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\right) \end{aligned} \tag{24}$$

$$\begin{aligned} l_{\boldsymbol{\beta}\gamma_s}(\boldsymbol{\theta}) &= -\sum_{i=1}^{n}\nabla_{\boldsymbol{\beta}}\mathbf{g}(\mathbf{x},\boldsymbol{\beta})\Omega'^{-1}\mathbf{e}_i\mathbf{e}_i'\frac{\partial\Omega^{-1}}{\partial\gamma_s}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\mathcal{H}_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\right) \\ &\quad - \sum_{i=1}^{n}\nabla_{\boldsymbol{\beta}}\mathbf{g}(\mathbf{x},\boldsymbol{\beta})\frac{\partial\Omega'^{-1}}{\partial\gamma_s}\mathbf{e}_i\mathcal{S}_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\right), \quad s=1,\ldots,m \end{aligned} \tag{25}$$

$$\begin{aligned} l_{\gamma_s\gamma_s}(\boldsymbol{\theta}) &= -\mathrm{tr}\left(\frac{\partial\Omega^{-1}}{\partial\gamma_s}\frac{\partial\Omega}{\partial\gamma_s} + \Omega^{-1}\frac{\partial^2\Omega}{\partial\gamma_s^2}\right) + \sum_{i=1}^{n}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]'\frac{\partial\Omega'^{-1}}{\partial\gamma_s}\mathbf{e}_i\mathbf{e}_i'\frac{\partial\Omega^{-1}}{\partial\gamma_s}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\mathcal{H}_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\right) \\ &\quad + \sum_{i=1}^{n}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]'\frac{\partial^2\Omega^{-1}}{\partial\gamma_s^2}\mathbf{e}_i\mathcal{S}_i\left(\mathbf{e}_i'\Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})]\right), \quad s=1,\ldots,m \end{aligned} \tag{26}$$

where $\mathcal{H}_i(\epsilon) = \frac{d^2l_i(\epsilon)}{d\epsilon^2}$ is the second-order derivative of the individual likelihood and $\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}'}\mathbf{g}(\mathbf{x},\boldsymbol{\beta})$ is the $k\times k$ matrix of second order partial derivatives of $\mathbf{g}(\mathbf{x},\boldsymbol{\beta})$. From this, we calculate $\hat{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'} = j_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}})$ and $\hat{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'} = j_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\hat{\boldsymbol{\theta}}_k)$ using the full and constrained maximum likelihood values.

For the calculation of $Q$, we need the conditioning vectors $\mathbf{V}$ in (15) obtained from an appropriate pivotal quantity. We consider the location-scale standardized coordinates in the vector quantity

$$\mathbf{q}(\mathbf{y},\mathbf{x};\boldsymbol{\theta}) = \Omega^{-1}[\mathbf{y}-\mathbf{g}(\mathbf{x},\boldsymbol{\beta})],$$

which is an $n$-dimensional quantity with a $\boldsymbol{\theta}$-free distribution. The tangent directions for an ancillary are then obtained using formula (15) as the following $n \times (k+m)$ matrix.

$$\mathbf{V} \quad = \quad \Big( \; \nabla_{\boldsymbol{\beta}'} \mathbf{g}(\mathbf{x}, \hat{\boldsymbol{\beta}}), \quad \hat{\Omega}^{-1}_{\gamma_1} \hat{\Omega}^{-1} [\mathbf{y} - \mathbf{g}(\mathbf{x}, \hat{\boldsymbol{\beta}})] \quad , \ldots, \quad \hat{\Omega}^{-1}_{\gamma_m} \hat{\Omega}^{-1} [\mathbf{y} - \mathbf{g}(\mathbf{x}, \hat{\boldsymbol{\beta}})] \; \Big)$$

where $\hat{\Omega}^{-1} = \Omega^{-1}(\hat{\boldsymbol{\gamma}})$ and $\hat{\Omega}^{-1}_{\gamma_s} = \left. \dfrac{\partial \Omega^{-1}(\boldsymbol{\gamma})}{\partial \gamma_s} \right|_{\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}}$, $\quad s = 1, \ldots, m$. These vectors $\mathbf{V}$ are then used in formula (14) to obtain the nominal exponential parameterization, $\boldsymbol{\varphi}(\boldsymbol{\theta})$. The values of $\chi(\boldsymbol{\theta})$ and $Q$ follow by direct calculations from (10) to (13). Highly accurate tail probabilities can then be obtained by either the Lugannani and Rice (LR) formula or the Barndorff-Nielsen (BN) formula given by (16) and (17) respectively.

These results are now accessed in numerical studies to demonstrate their finite sample performances in three simulations and a real data example.

# 5 Monte Carlo Simulations

The goal of the simulation study is to assess the small sample performance of the proposed methods and compare them with other methods. The accuracy of the different methods is evaluated by computing central coverage and tail probabilities at the nominal 95% confidence interval for the parameters of interest. In particular, we compute the proportion of the true parameter value falling within the method's confidence interval (central coverage), the proportion of the true parameter value falling above the upper limit of the method's confidence interval (upper error), and the proportion of the true parameter value falling below the lower limit of the method's confidence interval (lower error). The labels $r$, Wald, $LR$, and $BN$ refer to the signed log-likelihood ratio method, the Wald method, the Lugannani and Rice method and the Barndorff-Nielsen method, given by formulas (9), (10), (16) and (17), respectively.

## 5.1 Simulation Study 1: Linear regression with non-normal errors

Consider the linear regression model defined by

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n,$$

with $\beta_0 = 1$ and $\beta_1 \in \{0, 1\}$. The explanatory variable is log-normally distributed i.e. $\log(x_i) \sim N(0, 1)$.

*Case 1: The errors follow a student distribution:* $u_i = \sigma \epsilon_i$, *with* $\epsilon_i \sim t_{(7)}$ *and* $\sigma \in \{1, 2\}$

This model corresponds to the one given in (18) with $\mathbf{g}(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ and $\Omega(\sigma) = \sigma \mathbf{I}_n$, where $\boldsymbol{\beta} = [\beta_0 \;\; \beta_1]'$, $\mathbf{x} = [x_1 \ldots x_n]'$, $\mathbf{X} = [\mathbf{i} \;\; \mathbf{x}]$ with $\mathbf{i} = [1 \ldots 1]'$, and $\boldsymbol{\epsilon}$ is a vector of independent Student random variables with 7 degrees of freedom. With $N = 100,000$ replications of sample size $n = 5$ from the Student distribution, we calculate the corresponding $p$-values for testing $\beta_1 = 0$, and $\sigma = 1$, using approximations (16) and (17). The target accuracy is the uniform $(0, 1)$ distribution.

The results in percentages are recorded in Table 1. The Nominal values are the targeted (or desired) sizes for 95% confidence interval inference. The ability of the third-order methods (BN and LR) to give close

Table 1: **Results of Simulation study 1 for 95% CI**

| Hypothesis | Method | Upper Error | Lower Error | Central Coverage |
|---|---|---|---|---|
| $\beta_1 = 0$ | Nominal | 0.0250 | 0.0250 | 0.9500 |
| | BN | 0.0261 | 0.0242 | 0.9497 |
| | LR | 0.0258 | 0.0253 | 0.9489 |
| | $r$ | 0.0693 | 0.0176 | 0.9131 |
| | Wald | 0.0688 | 0.0154 | 0.9158 |
| $\sigma = 1$ | Nominal | 0.0250 | 0.0250 | 0.9500 |
| | BN | 0.0265 | 0.0256 | 0.9479 |
| | LR | 0.0261 | 0.0258 | 0.9481 |
| | $r$ | 0.0657 | 0.0217 | 0.9126 |
| | Wald | 0.0634 | 0.0258 | 0.9108 |

approximations when $n = 5$ attests to their reliability and superior accuracy in comparison with the classical tests when the sample size is small. In particular, while the third-order methods produce upper and lower error probabilities that are relatively symmetric, with a tail probability totaling approximately 5%, those produced by the first-order methods are heavily skewed, with higher total error rates.

## 5.2 Simulation Study 2: Nonlinear regression with homoskedastic errors

Consider the nonlinear regression model defined by

$$y_i = \beta_0[1 + \exp(-x_i\beta_1)]^{-1} + \sigma\epsilon_i$$

where the errors $\epsilon_i$ are either normal or Student(6), but the researcher does not know which of the two families the errors belong to. We set $\beta_0 = \beta_1 = 1$, $\sigma = 1$, and examine a small sample of size $n = 7$ and $x_i = 0, 1, 2, 4, 7, 10, 12$.

We first simulate a normal model (i.e. $\epsilon_i \sim N(0, 1)$) and perform the analysis on the basis of the correct (normal) model and incorrect (Student) model. Likewise, we simulate a Student (6) model (i.e $\epsilon_i \sim t_6$) and perform the analysis on the basis of the correct (Student) model and incorrect (Normal) model. We perform $N = 100,000$ replications for each case. The results obtained from testing the hypothesis $\beta_1 = 1$ are presented in Table 2. We note that when the model errors are correctly specified the third-order methods LR and BN give highly accurate results that are very close to the targeted nominal frequency while first-order methods, $r$ and Wald are worse. The characteristics of the tail probabilities are similar to the previous example, i.e., the proposed approaches have symmetric tail probabilities close to their nominal targets, while the classical methods have highly skewed tail probabilities and are large in total error rates. When the errors are incorrectly specified, the accuracy of the proposed third-order methods, BN and LR deteriorate, but still give reasonable error rates while the classical methods are much worse, with total error rates sometimes exceeding 18%.

Table 2: **Results of Simulation study 2 for 95% CI**

| Model | Method | Upper Error | Lower Error | Central Coverage |
|---|---|---|---|---|
| Normal | | *Normal Analysis* | | |
| | Nominal | 0.0250 | 0.0250 | 0.9500 |
| | BN | 0.0259 | 0.0254 | 0.9487 |
| | LR | 0.0261 | 0.0255 | 0.9484 |
| | $r$ | 0.0689 | 0.0297 | 0.9014 |
| | Wald | 0.0601 | 0.0368 | 0.9031 |
| | | *Student Analysis* | | |
| | Nominal | 0.0250 | 0.0250 | 0.9500 |
| | BN | 0.0405 | 0.0316 | 0.9279 |
| | LR | 0.0352 | 0.0331 | 0.9317 |
| | $r$ | 0.0971 | 0.0711 | 0.8318 |
| | Wald | 0.0963 | 0.0802 | 0.8235 |
| Student | | *Normal Analysis* | | |
| | Nominal | 0.0250 | 0.0250 | 0.9500 |
| | BN | 0.0371 | 0.0351 | 0.9278 |
| | LR | 0.0395 | 0.0359 | 0.9246 |
| | $r$ | 0.0717 | 0.0979 | 0.8304 |
| | Wald | 0.0798 | 0.0958 | 0.8244 |
| | | *Student Analysis* | | |
| | Nominal | 0.0250 | 0.0250 | 0.9500 |
| | BN | 0.0261 | 0.0259 | 0.9480 |
| | LR | 0.0263 | 0.0258 | 0.9479 |
| | $r$ | 0.0377 | 0.0507 | 0.9116 |
| | Wald | 0.0335 | 0.0579 | 0.9086 |

## 5.3 Simulation Study 3: Linear regression with heteroskedastic errors

In this simulation we specify a linear model with heteroskedastic errors similar to Luger (2010)

$$y_i = x_i'\beta + \exp(z_i\gamma)\epsilon_i$$

where $\epsilon_i$ is i.i.d. according to a $N(0,1)$ distribution. Here $x_i$ is a $2 \times 1$ vector with the first element equal to one and the other element is a standard normal random variable. Likewise, $z_i$ is an independent standard nornal random variable. The regression parameters $\beta$ are set equal to a vector of ones, and the scalar $\gamma$ takes values in $[0,1]$. When $\gamma = 0$, the model is homoskedastic and when $\gamma > 0$ the model is heteroskedastic. We examine a small sample of size $n = 10$ and we perform inference using $N = 100,000$ replications for each of the case where $\gamma = 0$ and $\gamma = 1$. The results obtained from testing these hypotheses are reported in Table 3.

As for the previous cases, this example confirms that while the third-order methods produce upper and lower error probabilities that are relatively symmetric, with a tail probability totaling approximately 5% as targeted and central coverage close to the nominal level of 95%, those produced by the classical first-order methods are sometimes skewed with the total error probability as high as 11%.

Table 3: **Results of Simulation study 3 for 95% CI**

| Hypothesis | Method | Upper Error | Lower Error | Central Coverage |
|---|---|---|---|---|
| $\gamma = 0$ | Nominal | 0.0250 | 0.0250 | 0.9500 |
| | BN | 0.0263 | 0.0261 | 0.9476 |
| | LR | 0.0259 | 0.0260 | 0.9481 |
| | $r$ | 0.0619 | 0.0471 | 0.8910 |
| | Wald | 0.0643 | 0.0448 | 0.8909 |
| $\gamma = 1$ | Nominal | 0.0250 | 0.0250 | 0.9500 |
| | BN | 0.0261 | 0.0258 | 0.9481 |
| | LR | 0.0262 | 0.0261 | 0.9477 |
| | $r$ | 0.0692 | 0.0401 | 0.8907 |
| | Wald | 0.0641 | 0.0492 | 0.8867 |

Unreported simulations show that as we increase the sample size $n$, all the methods tend to be approximately similar, with central coverage close to 95% while tail probabilities approach 5%. However, the first-order methods converge much slower, as shown by the theory.[4]

# 6 Empirical Application: Mobile Money "M-Pesa" in Kenya

The use of mobile banking through cell phones plays an important role in the growth of economies, particularly in less industrialized countries that generally have little access to formal financial services. A pioneering effort in Kenya, "M-Pesa" (*M* for mobile, *Pesa* for money in Swahili language) has grown to be the world's largest mobile money service. The data for this application are taken from Jack and Suri (2014) to fit a regression model of mobile money adoption rate in Kenya and perform parameter inference using the procedures discribed above. We also compare the results with those from traditional methods. The data that we use are quarterly data ranging from 2007:1 to 2011:1 reporting the number of mobile money users over time (that is, from april 2007 to april 2011). This application also perfectly suits our framework because, given that the mobile money technology was relatively recent in Kenya in this period, only a small number of time observations ($n = 13$) was available.

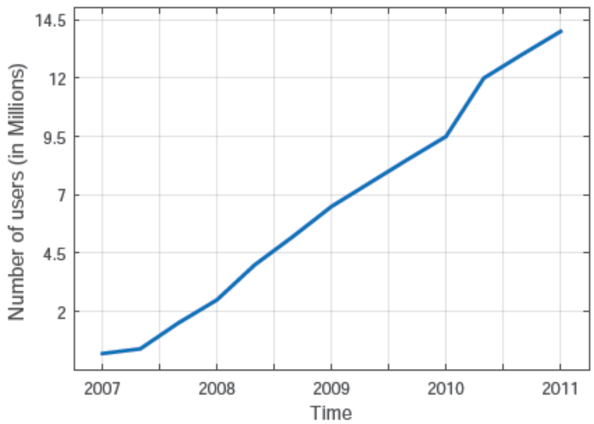
Figure 1(a) shows that M-PESA has grown rapidly since its launch in 2007 to reach 14 million registered users in 2011. This corresponds to about 70% of the adult population in Kenya. On the other hand, Figure 1(b) shows that this growth has occured at a decreasing rate tending to zero with time. The mobile money adoption process in Kenya thus exhibits the standard features of technology diffusion processes. They are expected to have an "S-shape" over time, which is typical to processes that consist of a slow early adoption stage, followed by a phase of rapid adoption which then tails off as the adopting population becomes saturated. A common approach used to model these processes is the Logistic trend curve defined by[5]

$$y_i = \beta_1 \left[1 + \beta_2 \exp(-\beta_3 x_i)\right]^{-1} + \sigma_i \varepsilon_i,$$

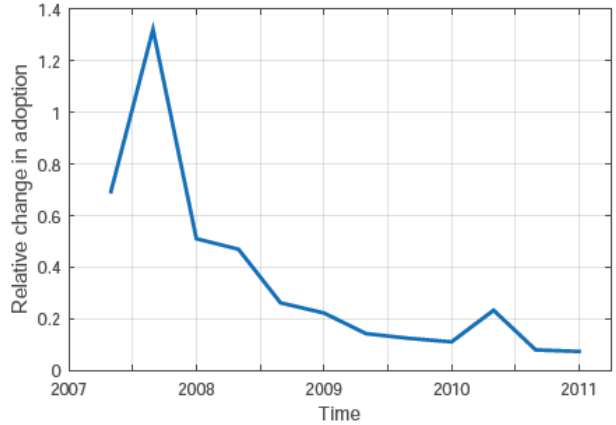
[4] These additional simulations results are available from the author.

[5] An overview of the so-called growth curves is given in Mahajan, Muller and Bass (1993); see also Meade and Islam (1995).

Figure 1: Levels and relative changes in mobile money adoption




(a) Number of mobile money users




(b) Relative change in mobile money users

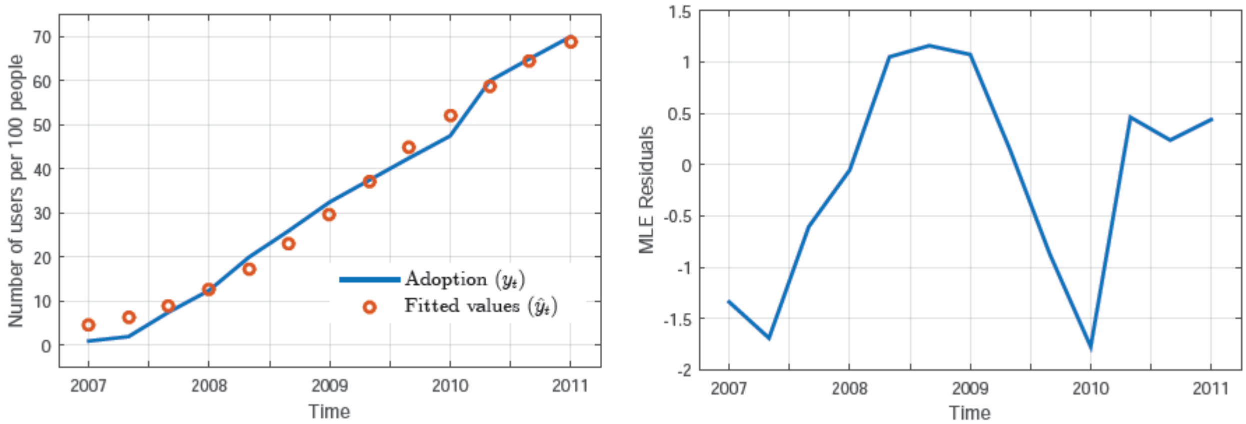
where $y_i$ is the number of mobile money users per 100 people at time $x_i = 1, 2, \ldots, n$, $\varepsilon_i \sim N(0, 1)$, and $g(x_i, \beta) = \beta_1 [1 + \beta_2 \exp(-\beta_3 x_i)]^{-1}$ is the regression function, with $\beta_1, \beta_2, \beta_3 > 0$. As for the conditional response variance, we allow it to depend nonlinearly on the time covariate as in Davison and Hinkley (1997) by $\sigma_i^2 = \sigma^2(1 + x_i)^{\gamma}$, where the two variance parameters $\gamma$ and $\sigma$ are to be estimated. Notice that when $\gamma = 0$, the model's disturbances are homoscedastic.[6]

The parameter $\beta_1$ represents the carrying capacity, i.e. the maximum size that can be reached with the available resources (e.g. level of mobile phone penetration), the parameter $\beta_2$ reflects the relative displacement and $\beta_3$ is the growth rate of mobile money usage. Figure 2(a) plots mobile money usage (in terms of the number of mobile money users per 100 people in the adult population) over the sample period in Kenya as well as the nonlinear regression fit of the data using the Logistic regression function, which appears to be graphically satisfactory. The estimation of the model was performed using numerical optimization that requires reasonable starting values for the parameters. There are no particular rules for starting values except that they should be as meaningful as possible to be close enough to the final values. I assume a starting value of 100 for $\beta_1$, that is, the number of users will reach at least 100 per 100 people in the adult population. If we scale time so that $x_0 = 0$ corresponds to April 2007, then using $y_0 = 1.0$ from the data and substituting $\beta_1^0 = 100$ in the model assuming zero error yields $\beta_2^0 = 99$. Returning to the data, at time $x_1 = 1$, the number of users per 100 adult people was $y_1 = 2.0$. Using this value along with previously determined start values and again setting the error to 0 gives $\beta_3^0 = 0.88$. Estimation results using maximum likelihood estimation (MLE) are given in Table 4.

The relative growth rate is estimated at $\hat{\beta}_3 = 0.38$ and the displacement parameter is estimated at $\hat{\beta}_2 = 24.75$. The carrying capacity is estimated at $\hat{\beta}_1 = 81.02$. This means that in the long run the number

[6]The usual common competitor for the Logistic growth model when modeling technology adoption is the Gompertz growth model. But preliminary analysis using the Nguimkeu (2014) selection test allows to retain the former for these data.

Figure 2: Adoption rate, fitted values, and MLE residuals




(a) Adoption rate and fit

(b) MLE residuals

Table 4: ML estimation results

| Parameters | ML Estimates | Standard Errors |
|---|---|---|
| $\beta_1$ | 81.025 | 6.4478 |
| $\beta_2$ | 24.748 | 4.4278 |
| $\beta_3$ | 0.3801 | 0.0382 |
| $\gamma$ | -1.359 | 0.8636 |
| $\ln\sigma$ | 0.9670 | 0.2773 |
| Log Likelihood | - 31.02 | |
| No of obs. | 13 | |

of mobile money users per 100 adult people would be about 81.

The algorithm described in Section 4 can therefore be applied to provide $p$-value functions for the parameters $\beta_1, \beta_2, \beta_3$, $\gamma$, and $\sigma$ as well as corresponding confidence intervals for each of the methods (BN, LR, Wald, $r$). Table 5 reports the 95% confidence interval for the model parameters from each of the methods. Each of these methods agree that all the regression coefficients of the model are statistically significant at the 5% level, except the variance parameter $\gamma$ which is insignificant, pointing to homoskedasticity. However, the confidence intervals obtained in Table 5 show considerable differences that could lead to different inferences about the parameters. While the third-order methods, BN and LR give very similar confidence intervals, the first-order methods, $r$ and Wald, on the other hand, give very different intervals. For example, a value of $\gamma = 1$ can not be statistically distinguished from 0 in the higher-order inference, whereas this value is statistically different from 0 in the first-order inference. Theoretically, the former are more accurate.

Table 5: 95% Confidence Intervals for Parameter Inference

| Parameters | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\gamma$ | $\ln\sigma$ |
|---|---|---|---|---|---|
| BN | [69.70 114.4] | [16.51 42.30] | [0.2959 0.4841] | [-3.016 1.2193] | [0.7335 1.6445] |
| LR | [68.84 106.8] | 16.31 41.40] | [0.2896 0.4833] | [-3.043 1.0789] | [0.7335 1.6445] |
| $r$ | [70.39 100.7] | [17.69 38.22] | [ 0.3041 0.4672] | [-3.079 0.5146] | [0.6265 1.4075] |
| Wald | [68.39 93.66] | [16.07 33.43] | [0.3051 0.4550] | [-3.051 0.3337] | [0.5825 1.3515] |
| Boot | [68.90 99.65] | [16.12 43.35] | [0.3020 0.4701] | [-3.018 1.3724] | [0.7502 1.5164] |

# 7 Conclusion

Recently developed third-order likelihood theory is used to propose methods to obtain highly accurate p-values for the parameters of regression models with small samples, by combining classical first-order approximation methods. Confirming the theory, simulation results show that significantly improved inferences can be made by using these third-order likelihood methods compared to the classical likelihood-based methods (i.e., signed log-likelihood ratio statistic, Wald statistic). The proposed methods are found to outperform the classical test statistics in every case including nonlinearity, non-normality and heteroskedasticity, and across all criteria considered such as tail error probabilities, central coverage and average bias. An empirical example with a logistic model of technology adoption using data from mobile money ("M-pesa") in Kenya is used to illustrate the usefulness of the proposed methods. Since these methods can be easily computationally implemented, rely on familiar likelihood-based quantities and are highly tractable they are viable alternatives to conventional methods in Econometrics. Extensions to these methods may include the consideration of improved inference for linear (or nonlinear) combinations of model parameters, or the consideration of non-regression models such as the Generalized Method of Moments (GMM), and other systems of equations that do not rely on finite sample distributional assumptions.

# References


Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70 343–365.

Barndorff-Nielsen, O.E. (1986). Inference on full and partial parameters based on the standardized signed log-likelihood ratio. *Biometrika*, 73, 307-322.

Barndorff-Nielsen, O.E. (1991). Modified signed log-likelihood ratio. *Biometrika,* 78, 557-563.

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

Brazzale, A. R., & Davison, A. C. (2008). Accurate Parametric Inference for Small Samples. *Statistical Science*, 23(4), 465-484.

Brazzale, A. R., Davison, A. C. & Reid, N. (2007). *Applied Asymptotics: Case Studies in Small Sample Statistics.* Cambridge Univ. Press. M

Butler, R. W. (2007). *Saddlepoint Approximations with Applications.* Cambridge Univ. Press.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25, 631–650.

Daniels, H. E. (1987). Tail probability approximations. *International Statistical Review*, 54, 37–48.

Davison, A. C. (2003). *Statistical Models.* Cambridge Univ. Press.

Davison, A. C., Fraser, D. A. S. and Reid, N. (2006). Improved likelihood inference for discrete data. *Journal of the Royal Statistical Society: Series B*, 68, 495–508.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge university press.

Elkantassi, S., Bellio, R., Brazzale, A. R., & Davison, A. C. (2023). Improved inference for a boundary parameter. *Canadian Journal of Statistics.*

Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, 77 65–76.

Fraser, D. A. S. (2017). p-values: The insight to modern statistical inference. *Annual Review of Statistics and its Application,* 4, 1–14.

Fraser, D. A. S., Andrews, D. A. & Wong, A. (2005). Computation of distribution functions from likelihood information near observed data. *Journal of Statistical Planning & Inference*, 134, 180–193. M

Fraser, D. & Reid, N., (1995). Ancillaries and Third Order Significance. *Utilitas Mathematica*, 47, 33-53.

Fraser, D. A. S., Reid, N. & Wu, J. (1999). A simple formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, 86 249–264.

Fraser, D., Reid, N., Li, R. & Wong, A. (2003). P-value formulas from likelihood asymptotics: bridging the singularities. *Journal of Statistical Research*, 37, 1-15.

Fraser, D. A. S., Rekkas, M., & Wong, A. (2005). Highly accurate likelihood analysis for the seemingly unrelated regression problem. *Journal of Econometrics*, 127(1), 17-33.

Jack, W., & Suri, T. (2014). Risk sharing and transactions costs: Evidence from Kenya's mobile money revolution. *American Economic Review*, 104(1), 183-223.

Lugannani, R. & Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probabity*, 12 475–490.

Luger, R. (2010). An omnibus test for heteroskedasticity. *Economics Letters,* 106(1), 22-24.

Mahajan, V., Muller, E., & Bass, F. M. (1993). New-product diffusion models. *Handbooks in operations research and management science,* 5, 349-408.

Meade, N., & Islam, T. (1995). Prediction intervals for growth curve forecasts. *Journal of Forecasting,* 14(5), 413-430.

Nguimkeu, P. (2014). Improved inference for moving average disturbances in nonlinear regression models. *Journal of Probability and Statistics*, 2014.

Nguimkeu, P. E., & Rekkas, M. (2011). Third-order inference for autocorrelation in nonlinear regression models. *Journal of Statistical Planning and Inference,* 141(11), 3413-3425.

Reid, N. (1988). Saddlepoint methods and statistical inference (with discussion). *Statistical Science*, 3, 213–238.

Reid, N. (1995). The roles of conditioning in inference (with discussion). *Statistical Science*, 10, 138–199.

Reid, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics*, 31, 1695–1731.

Severeni, T. (2000). *Likelihood Methods in Statistics.* Oxford University Press, New York.

Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, 2, 145-166

Skovgaard, I. M. (2001). Likelihood asymptotics. *Scandinavian Journal of Statistics*, 28, 3–32.