Numerical methods of sufficient sample size estimation for generalised linear models

A. V. Grabovoy,[1, *] **T. T. Gadaev,**[1, **]
A. P. Motrenko,[1, ***] **and V. V. Strijov**[1, ****]

Andrey Grabovoy

[1]*Moscow Institute of Physics and Technology 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russian Federation*

Received

Abstract—This paper investigates the problem of cost reduction of data collection procedures. To select an adequate model for regression or classification, a sample set of minimum sufficient size is required. This sample set must follow the data generation hypothesis: generalised linear regression models require independent and identically distributed target variables. Several numerical methods of sample size estimation are analysed and compared in practical terms. Statistic-based, heuristics and Bayesian types of methods are considered. Since the practical goal of the data collection is to construct a forecasting model, several tests use the model parameters. The computational experiment includes widely used sample sets. The open-source code and the software for the practitioners are provided.

Keywords and phrases: *Sample size estimation, Linear model, Bayesian approach, Statistics, Numerical methods*

1. INTRODUCTION

The design of experiment requires to estimate the minimum sample size: the quantity of the performed feature set measurements which are required to build the formulated conditions. The choice of the sample size estimation method depends on the problem being solved which determines the formulation of the statistical hypothesis and statistics to check it. Table 1 presents ten sample size estimation methods. It includes both classical and Bayesian methods of the sample size estimation.

The classical methods assume that the sample corresponds to some prior conditions formulated earlier. These conditions are formulated as a statistical criterion (Self et al. 1988, 1992; Shieh 2000; Demidenko 2007). The sample size estimation method related to this criterion guarantees that the fixed statistical power $1 - \beta$ with the extent of the first kind error which does not exceed the set value α will be approached. This sample size is called sufficient.

However, the practical applications of the sample size estimation methods assume a model to fit the measured data (Kloek 1975). These models are selected according to either regression or classification problem statement. In this paper generalised linear models are In the paper (Self et al. 1988), a method of power estimation of the Lagrange multiplier test for coefficients of the generalised linear regression, with the help of which the sample size is estimated, is described. The weakness of the method is in the fact that when an alternative hypothesis differs greatly from the null hypothesis, the maximum likelihood estimates for the model parameters and covariance matrix used for power rating are not asymptotically consistent in the alternative hypothesis. Later (Self et al. 1992) an approach to the estimation of power and sample size related to it was proposed on the

* E-mail: grabovoy.av@phystech.edu
** E-mail: gadaev.tt@phystech.edu
*** E-mail: anastasiya.motrenko@phystech.edu
**** E-mail: strijov@phystech.edu

Table 1. Methods

Method	Short overview	Reference
Lagrange multipliers test	Likelihood of the sample has the following form: $p(y\|\mathbf{x},\mathbf{w}) = \exp\big(y\theta - b(\theta) + c\,(y)\big)$. Sufficient sample size m^*: $m^* = \frac{\gamma^*}{\gamma^0}$, where γ^* and γ_0 can be found in (12) and (11).	Self et al. 1988
Likelihood ratio test	Likelihood of the sample has the following form: $p(y\|\mathbf{x},\mathbf{w}) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c\,(y,\phi)\right)$. Sufficient sample size m^*: $m^* = \frac{\gamma^*}{\Delta^*}$, where γ^* and Δ^* can be found from (17) and (19).	Shieh 2000
Wald statistic	Likelihood of the sample has the following form: $p(y\|\mathbf{x},\mathbf{w}) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c\,(y,\phi)\right)$. Sufficient sample size m^*: $m^* = \frac{\gamma^*}{\delta}$, where$\gamma^*$ and δ can be found from (26).	Shieh 2005
Cross-validation	Sufficient sample size m^*: $\forall m \geq m^*\ RS(m) \geq 1-\varepsilon$, where ε is chosen such that, RS is defined in (31).	Motrenko et al. 2014
Bootstrap	Sufficient sample size $m^* : \forall m \geq m^* \max_i (b_i^m - a_i^m) < l$, where (a_i^m, b_i^m) is quantile bootstrap confident interval calculated on i-th bootstrap subsample of size m	(Qumsiyeh 2013)
Kullback-Leibler	Sufficient sample size m^*: $\forall \mathfrak{D}_{\mathcal{B}_1} :\ \|\mathfrak{D}_{\mathcal{B}_1}\| \geq m^*\ \mathsf{E}_{\mathfrak{D}_{\mathcal{B}_2}} D_{\mathrm{KL}}\,(p_1, p_2) \leq \varepsilon$, where $\mathcal{B}_1$, $\mathcal{B}_2$ satisfy (42).	Motrenko et al. 2014
Average posterior variance criterion	Sufficient sample size m^*: $\forall m \geq m^*\ \mathsf{E}_{\mathfrak{D}_m} \mathsf{D}\,[\hat{\mathbf{w}}\|\mathfrak{D}_m] \leq \varepsilon$, where ε is sufficiently small.	Joseph et al. 1997, 1995
Average coverage criterion	Sufficient sample size m^*: $\forall m \geq m^*\ \mathsf{E}_{\mathfrak{D}_m} \mathsf{P}\,\{\mathbf{w} \in A\,(\mathfrak{D}_m)\} \geq 1-\alpha$, where α is sufficiently small.	Joseph et al. 1997, 1995
Average length criterion	Sufficient sample size m^*: $\forall m \geq m^*\ \mathsf{E}_{\mathfrak{D}_m} r_m \leq l$, where r_m is described in (39)	Joseph et al. 1997, 1995
Utility maximization	Sufficient sample size m^*: $m^* = \arg\max_m \mathsf{E}_{\mathfrak{D}_m} \int_{\mathbf{w}} u\,(\mathfrak{D},\mathbf{w})\, p(\mathbf{w}\|\mathfrak{D}) d\mathbf{w}$, where utility function $u\,(\mathfrak{D},\mathbf{w})$ is given as (41).	Lindley 1997

basis of the maximum likelihood ratio test. This approach appeared to be more accurate for a series of independent variables. Besides, a power estimation method for Wald statistics was proposed in the paper (Shieh 2005). In the paper (Motrenko et al. 2014) in case of logistic regression, it is proposed that the method which uses the ROC-AUC curve and shift concept be used. The classical methods (Self et al. 1988, 1992; Shieh 2000, 2005; Demidenko 2007) have a series of restrictions related to practical application of these methods. In order to estimate the sample size, it is required to know the parameter estimation variance or, in a more general case, to have the estimation of the non-centrality parameter in the distribution of the statistics used when the alternative hypothesis is true. These methods do not show how to obtain these values. Besides, the estimation variance and non-centrality parameter will not be obtained with a certain variance the influence of which on the sample size estimation result is irrelevant.

Statistical methods make it possible to estimate the sample size on the basis of assumptions about the distribution of data and information about the correspondence between the values observed

and the assumptions of the null hypothesis. When the size of the sample under investigation is sufficient or excessive, it is possible to use the methods based on the observation of alteration of certain characteristic of the model building procedure when enhancing the sample size. In particular, when observing the relation of the forecasting quality with the control sample and training sample (Motrenko et al. 2014), we shall determine the sufficient sample size which corresponds to the start of over-training. In the paper (Qumsiyeh 2013), a bootstrap method is used to estimate the sufficient sample size. The excess of the current sample size is checked on the basis of a confident intervals analysis of the parameter estimated. The width of the confident interval with different values of the sample size is estimated with the help of a bootstrap method. For this purpose, the samples of smaller size are sampled the specified number of times, and the confident interval for an error when estimating the model parameter is calculated. The sample size is considered sufficient when the width of the confident interval does not exceed a certain value set in advance.

The restrictions of statistical methods of sample size estimation listed above are considered in details in Bayesian procedure (Lindley 1997; Rubin et al. 1998; Wang et al. 2002) where the sample size estimation is determined on the basis of maximisation of the expected merit function (Lindley 1997). The merit function may include the explicit parameter distribution functions and penalties for the sample size enhancement. The alternative to the approaches (Wang et al. 2002) based on the merit function is the sampling of the sample size by setting restrictions on a certain model parameter estimation quality criterion. The examples of such criteria are the following: average posterior variance criterion (AVPC), average length criterion (ALC), average coverage criterion (ACC). For every criterion listed, the sample size estimation is determined as a minimum value of the sample size for which the expected value of the criterion chosen does not exceed any fixed threshold. In the paper (Motrenko et al. 2014), it is proposed that the sample size be considered sufficient if the space between the distributions estimated on the basis of subsamples of this size is sufficiently small. Such approach does not require any further generalisation in case of multiple variables. Besides, estimation may be made in the presence of data distribution assumptions, as well as in their absence. The weakness of this approach is in the fact that quantitative estimation can be obtained only when the sample size is excessive.

2. PROBLEM STATEMENT OF MODEL CONSTRUCTION

Given a sample set of size m:

$$\mathfrak{D}_m = \{\mathbf{x}_i, y_i\}_{i=1}^m \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{Y}$. Feature vector $\mathbf{x} = [\mathbf{u}, \mathbf{v}]$ concatenates $\mathbf{u}_i \in \mathbb{R}^k$ and $\mathbf{v}_i \in \mathbb{R}^{n-k}$. The sample set $\mathfrak{D}_m$ randomly splits into train and test parts

$$\mathfrak{D}_{\mathcal{T}_m} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i\in\mathcal{T}_m}, \quad \mathfrak{D}_{\mathcal{L}_m} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i\in\mathcal{L}_m}, \quad \mathcal{T}_m \sqcup \mathcal{L}_m = \{1, ..., m\}. \tag{2}$$

Let introduce a parametric family of functions for unknown distribution approximation $p(y|\mathbf{x}, \mathfrak{D}_{\mathcal{L}_m})$:

$$\mathfrak{F} = \left\{ f(y, \mathbf{x}, \mathbf{w}) \,|\mathbf{w} \in \mathbb{W}, \int_{y\in\mathbb{Y}, \mathbf{x}\in\mathbb{R}^n} f(y, \mathbf{x}, \mathbf{w})\, dy d\mathbf{x} = 1 \right\}. \tag{3}$$

For the model f with the parameters vector $\mathbf{w}$ define the likelihood function and logarithmic likelihood function of the sample set $\mathfrak{D}$:

$$L(\mathfrak{D}, \mathbf{w}) = \prod f(y, \mathbf{x}, \mathbf{w}), \quad l(\mathfrak{D}, \mathbf{w}) = \sum \log f(y, \mathbf{x}, \mathbf{w}), \tag{4}$$

where $f(y, \mathbf{x}, \mathbf{w})$ is the likelihood of the sample set $\mathfrak{D}_{\mathcal{L}}$ with given vector of parameters $\mathbf{w}$. Use maximum likelihood principle to estimate parameters $\mathbf{w}$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbb{W}} L(\mathfrak{D}_{\mathcal{L}}, \mathbf{w}). \tag{5}$$

The Fisher information matrix has the form:

$$\mathbf{I}(\mathfrak{D}, \mathbf{w}) = -\nabla\nabla^{\mathsf{T}} l(\mathfrak{D}, \mathbf{w}), \quad \mathbf{V} = \mathbf{I}^{-1}(\mathfrak{D}, \mathbf{m}), \tag{6}$$

statistic-based methods and Bayesian methods use the Fisher information matrix to estimation the sample size.

3. STATISTIC-BASED SAMPLE SIZE ESTIMATION

The main advantage of statistic-based methods is their capability of estimating sufficient sample size having an insufficient sample set. They allow to predict how many samples are needed on the early stage of experiment.

3.1. Lagrange multipliers test

Let the likelihood has the following form:

$$p(y|\mathbf{u},\mathbf{v},\mathbf{w}_u,\mathbf{w}_v)=\exp\big(y\theta-b(\theta)+c\,(y)\big), \tag{7}$$

where θ is a parameters of distribution, and it calculates by using link function $\theta=\theta\big(\mathbf{u},\mathbf{v},\mathbf{w}_u,\mathbf{w}_v\big)$.

Test the hypothesis

$$H_0:\mathbf{m}_u=\mathbf{m}_u^0,\quad H_1:\mathbf{m}_u\neq\mathbf{m}_u^0. \tag{8}$$

Let statistics $S_{m,u}\,(\mathbf{w}_u,\mathbf{w}_v)$ and $S_{m,v}\,(\mathbf{w}_u,\mathbf{w}_v)$ are derivatives of log-likelihood of the sample set $\mathfrak{D}_m$ with respect to $\mathbf{w}_u$ and $\mathbf{w}_v$. Consider $\mathbf{s}_m=S_{m,u}\left(\mathbf{m}_u^0,\hat{\mathbf{w}}_v^0\right)$, where $\hat{\mathbf{w}}_v^0$ is derived from the equation

$$S_{m,v}\left(\mathbf{m}_u^0,\mathbf{w}_v\right)=0. \tag{9}$$

Then the Lagrange statistic is

$$LM=\mathbf{s}_m^{\mathsf{T}}\mathbf{Q}_m^{-1}\mathbf{s}_m. \tag{10}$$

where $\mathbf{Q}_m$ is the covariance matrix of vector $\mathbf{s}_m$.

When H_0 holds, the statistic LM asymptotically follows a $\chi^2(k)$ distribution. In (Self et al. 1988) it is shown, that when an alternative hypothesis H_1 holds, LM asymptotically follows a distribution $\chi^2(k,\gamma)$, where γ is a non-centrality parameter

$$\gamma=\boldsymbol{\xi}_m^{\mathsf{T}}\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\xi}_m=m\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}=m\gamma^0, \tag{11}$$

where $\boldsymbol{\xi}_m$ and $\boldsymbol{\Sigma}_m$ are expectation and covariance matrix of $\mathbf{s}_m$. Denote $\boldsymbol{\xi}_1=\boldsymbol{\xi}$, $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}$.

The alternative method to derive γ involves the conditions on the significance level α and the probability of II type error β:

$$\gamma^*:\chi^2_{k,1-\alpha}=\chi^2_{k,\beta}\,(\gamma)\,. \tag{12}$$

Using (11) and (12) derive

$$m^*=\frac{\gamma^*}{\gamma^0}. \tag{13}$$

This is a sufficient minimum sample size to distinguish $\mathbf{m}_u$ from $\mathbf{m}_u^0$.

3.2. Likelihood ratio test

Let the likelihood of the sample be

$$p(y|\mathbf{u},\mathbf{v},\mathbf{w}_u,\mathbf{w}_v)=\exp\left(\frac{y\theta-b(\theta)}{a(\phi)}+c\,(y,\phi)\right), \tag{14}$$

where θ is a parameters of distribution, and it calculates by using link function $\theta=\theta\big(\mathbf{u},\mathbf{v},\mathbf{w}_u,\mathbf{w}_v\big)$.

Test the hypothesis:

$$H_0:\mathbf{m}_u=\mathbf{m}_u^0,\quad H_1:\mathbf{m}_u\neq\mathbf{m}_u^0. \tag{15}$$

Introduce the logarithm of likelihood ratio statistics:

$$LR=2\Big(l\,(\mathfrak{D},\hat{\mathbf{w}})-l\,\left(\mathfrak{D},\hat{\mathbf{w}}^0\right)\Big), \tag{16}$$

where $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ is the vector, which maximizes likelihood (14), $\hat{\mathbf{w}}^0 = [\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0]$ is the vector, which maximizes likelihood (14) with fixed $\mathbf{m}_u^0$.

When H_0 holds, the statistics LR asymptotically has $\chi^2(k)$ distribution. In (Shieh 2000) it is shown, that if the alternative hypothesis H_1 holds, LR asymptotically has distribution $\chi^2(k,\gamma)$, where γ is a non-centrality parameter, which is given as

$$\gamma = m\Delta^*, \quad \Delta^* = \mathsf{E}\left[2a^{-1}(\phi)\left\{(\theta-\theta^*)\nabla b(\theta) - b(\theta) + b(\theta^*)\right\}\right], \tag{17}$$

where the parameters θ and θ^* are calculated according to the parameters $\mathbf{w} = [\mathbf{w}_u, \mathbf{w}_v]$ and $\mathbf{w}^* = [\mathbf{w}_u^0, \mathbf{w}_v^*]$ respectively. The parameters $\mathbf{w}_v^*$ are given as the solution of the equation:

$$\lim_{m\to\infty} m^{-1}\mathsf{E}\left(\frac{\partial l\left(\mathfrak{D}, \left[\mathbf{m}_u^0, \mathbf{w}_v\right]\right)}{\partial \mathbf{w}_v}\right) = 0. \tag{18}$$

Then with given α and β the sufficient sample size m^* is

$$m^* = \frac{\gamma^*}{\Delta^*}, \quad \gamma^* : \chi^2_{k,1-\alpha} = \chi^2_{k,\beta}(\gamma), \tag{19}$$

where $\chi^2_{k,1-\alpha}$, $\chi^2_{k,\beta}(\gamma^*)$ are the quantiles of the distributions χ^2_k and $\chi^2_k(\gamma^*)$.

3.3. Wald test

Let the likelihood of the sample be

$$p(y|\mathbf{u},\mathbf{v},\mathbf{w}_u,\mathbf{w}_v) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{20}$$

where θ is a parameters of distribution, and it calculates by using link function $\theta = \theta\big(\mathbf{u},\mathbf{v},\mathbf{w}_u,\mathbf{w}_v\big)$.

Test the hypothesis:

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad H_1 : \mathbf{m}_u \neq \mathbf{m}_u^0. \tag{21}$$

The Wald test for the hypothesis is

$$W = \left(\hat{\mathbf{w}}_u - \mathbf{m}_u^0\right)^{\mathsf{T}} \hat{\mathbf{V}}_u^{-1} \left(\hat{\mathbf{w}}_u - \mathbf{m}_u^0\right), \tag{22}$$

where $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_u, \hat{\mathbf{w}}_v]$ is the vector of parameters, which maximizes likelihood (20), and $\hat{\mathbf{V}}_u$ is defined in (6).

If H_0 holds, the statistic W asymptotically has χ^2 distribution. In (Shieh 2005) it is shown that in case of H_1, the statistic W asymptotically follows a $\chi^2(k,\gamma)$ distribution, where γ is a noncentrality parameter:

$$\gamma = m\delta, \quad \delta = \left(\hat{\mathbf{w}}_u - \mathbf{m}_u^0\right)^{\mathsf{T}} \boldsymbol{\Sigma}_u^{-1} \left(\hat{\mathbf{w}}_u - \mathbf{m}_u^0\right), \quad \boldsymbol{\Sigma}_u = m\hat{\mathbf{V}}_u. \tag{23}$$

Using some given significance level α and the probability of type II error β, define the sample size estimation as

$$m^* = \frac{\gamma^*}{\delta}, \quad \gamma^* : \chi^2_{k,1-\alpha^*} = \chi^2_{k,\beta}(\gamma), \tag{24}$$

where $\chi^2_{k,1-\alpha^*}$, $\chi^2_{k,\beta}(\gamma^*)$ are quantiles of distributions and α^* is a correction on the significance levels:

$$\alpha^* = P\left(\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\xi} > \chi^2_{k,1-\alpha}\right), \quad \Sigma^* = \mathbf{I}^{-1}(\mathfrak{D},\mathbf{w}^*), \tag{25}$$

where $\mathbf{w}^* = \left[\mathbf{m}_u^0, \mathbf{w}_v^*\right]$ is a solution of the equation:

$$\lim_{m\to\infty} m^{-1}\mathsf{E}\left(\frac{\partial l\left(\mathfrak{D}, \left[\mathbf{m}_u^0, \mathbf{w}_v\right]\right)}{\partial \mathbf{w}_v}\right) = 0. \tag{26}$$

4. HEURISTICS-BASED SAMPLE SIZE ESTIMATION

The heuristics-based method uses popular statistical heuristics such as bootstrap, cross-validation and feature selection.

4.1. Sample size estimation for logistic regression

Introduce the set of indexes $\mathcal{A}$ for the logistic regression parameters $\mathbf{w}$. Test the hypothesis:

$$H_0 : j \notin \mathcal{A} \ \left(\mathrm{w}_j = 0\right), \quad H_1 : j \in \mathcal{A}^* \ \left(\mathrm{w}_j \neq 0\right), \tag{27}$$

where w_j is the jth element of the vector $\mathbf{w}$. When H_0 is not rejected, the vector $\mathbf{w}_{\mathcal{A}}$ holds. Set the margin c_0 for the logistic regression problem and obtain:

$$H_0 : 1 - c_0 = p_0, \quad H_1 : 1 - c_0 = p_1, \tag{28}$$

where c_0 is an optimal solution of the problem, when the feature j is excluded.

Use the statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{m}, \quad \hat{p} = \frac{1}{m}\sum_{i=1}^{m} y_i. \tag{29}$$

When H_0 is true, statistic Z is asymptotically distributed as $\mathcal{N}(0,1)$. In case of H_1, Z is asymptotically distributed as $\mathcal{N}\left(p_1 - p_0, \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)$.

The sufficient sample size is

$$m^* = \frac{p_0(1-p_0)\left(Z_{1-\alpha/2} + Z_{1-\beta}\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)^2}{(p_1 - p_0)^2}, \tag{30}$$

where $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are quantiles of $\mathcal{N}(0,1)$.

We will not consider this method further, since it can only be used for the logistic regression problem.

4.2. Cross-validation based method

Define the over-fitting criterion as

$$RS(m) = \ln \frac{L(\mathfrak{D}_{\mathcal{L}(m)}, \hat{\mathbf{w}})}{L(\mathfrak{D}_{\mathcal{T}(m)}, \hat{\mathbf{w}})}, \quad \frac{|\mathcal{T}(m)|}{|\mathcal{L}(m)|} = \mathrm{const} \leq 0.5. \tag{31}$$

Note that

$$\lim_{m\to\infty} RS(m) \to 0. \tag{32}$$

The sufficient sample size m^* is defined according to the condition:

$$m^* : \ \forall m \geq m^* \ \mathsf{E}_{\mathfrak{D}_m} RS(m) \leq \varepsilon, \tag{33}$$

where ε is an arbitrary parameter.

4.3. Bootstrap-based method

This method assumes that the lengths of the bootstrap quantile confident intervals do not exceed some fixed value l. Given some sample size m calculate the quantile confident intervals $(a_1^m, b_1^m), (a_2^m, b_2^m), ..., (a_n^m, b_n^m)$ with significance level of α using bootstrap for every parameter of the model. The sufficient sample size is

$$m^* : \forall m \geq m^* \max_i (b_i^m - a_i^m) < l. \tag{34}$$

Note that this method is coordinate-wise. Therefore, to increase the prediction accuracy is required a significant increase in the sample size.

5. BAYESIAN SAMPLE SIZE ESTIMATION

5.1. Model effectiveness analysis

The Bayesian methods of sample size estimation are based on a restriction of some model characteristics. For effectiveness analysis the function of sample size is defined. Increasing of this function is interpreted as decreasing of model effectiveness. Sample size m^* is chosen such that the explored function is lesser than some threshold ε.

a. Average posterior variance criterion. The sample size m^* is defined by the condition:

$$\forall m \geq m^* \ \mathsf{E}_{\mathfrak{D}_m} \mathsf{D}\left[\hat{\mathbf{w}} | \mathfrak{D}_m\right] \leq l. \tag{35}$$

where l is some given parameter, which quantifies the uncertainty of parameter estimation.

b. Average coverage criterion. Denote by $A\left(\mathfrak{D}\right) \subset \mathbb{R}^n$ be some set of the model parameters $\mathbf{w}$:

$$A\left(\mathfrak{D}\right) = \left\{\mathbf{w} : ||\mathbf{w} - \hat{\mathbf{w}}|| \leq l\right\}, \tag{36}$$

where l is some fixed ball radius. The sample size m^* is defined by the average coverage criterion:

$$\forall m \geq m^* \ \mathsf{E}_{\mathfrak{D}_m} \mathsf{P}\left\{\mathbf{w} \in A\left(\mathfrak{D}_m\right)\right\} \geq 1 - \alpha, \tag{37}$$

where α is some small value.

c. Average length criterion. Use the coverage of the model parameters $\mathbf{w}$ and define $A\left(\mathfrak{D}\right)$ as

$$\mathsf{P}\left(A\left(\mathfrak{D}\right)\right) = 1 - \alpha. \tag{38}$$

The average length criterion estimates m^* as in (37):

$$\forall m \geq m^* \ \mathsf{E}_{\mathfrak{D}_m} r_m \leq l, \tag{39}$$

where r_m is the ball radius $A\left(\mathfrak{D}_m\right)$.

5.2. Utility maximization

Methods of this class maximize the expectation of some utility function $u\left(\mathfrak{D}, \mathbf{w}\right)$ across the sample size:

$$m^* = \arg\max_m \mathsf{E}_{\mathfrak{D}_m} \int_{\mathbf{w}} u\left(\mathfrak{D}_m, \mathbf{w}\right) p(\mathbf{w}|\mathfrak{D}_m) d\mathbf{w}, \tag{40}$$

where utility function $u\left(\mathfrak{D}, \mathbf{w}\right)$ has the form:

$$u\left(\mathfrak{D}_m, \mathbf{w}\right) = l\left(\mathfrak{D}_m, \mathbf{w}\right) - cm, \tag{41}$$

where c is a penalization function for each element in the sample set.

5.3. Kullback-Leibler divergence

Call the index sets $\mathcal{B}_1, \mathcal{B}_2 \subset \{1, ..., m\}$ in the neighbourhood, if

$$|\mathcal{B}_1 \Delta \mathcal{B}_2| = 1. \tag{42}$$

So that $\mathcal{B}_2$ can be transformed into $\mathcal{B}_1$ by removal, replacement or addition of one element. In (Motrenko et al. 2014) it is shown that if the size of the sample set $\mathfrak{D}_{\mathcal{B}_1}$ is large enough, than the model parameters $\hat{\mathbf{w}}_1$, which are optimised with $\mathfrak{D}_{\mathcal{B}_1}$ must be in the neighbourhood of the model parameters $\hat{\mathbf{w}}_2$, which are optimised with $\mathfrak{D}_{\mathcal{B}_2}$.

Use Kullback-Leibler divergence as a proximity function between distributions of the model parameters, optimised with $\mathfrak{D}_{\mathcal{B}_1}$ and $\mathfrak{D}_{\mathcal{B}_2}$:

$$D_{\mathrm{KL}}\left(p_1, p_2\right) = \int_{\mathbf{w} \in \mathbb{W}} p_1(\mathbf{w}) \log \frac{p_1(\mathbf{w})}{p_2(\mathbf{w})} d\mathbf{w}, \tag{43}$$

where p_1 and p_2 are posterior probabilities of vector of parameters $\mathbf{w}$ calculated on subsamples $\mathfrak{D}_{\mathcal{B}_1}$ and $\mathfrak{D}_{\mathcal{B}_2}$ respectively. It is also assumed that $\mathfrak{D}_{\mathcal{B}_1}$ and $\mathfrak{D}_{\mathcal{B}_2}$ are in the neighbourhood. Then estimate the sample size m^* as:

$$\forall \mathfrak{D}_{\mathcal{B}_1} : \ |\mathfrak{D}_{\mathcal{B}_1}| \geq m^* \ \mathsf{E}_{\mathfrak{D}_{\mathcal{B}_2}} D_{KL}\left(p_1, p_2\right) \leq \varepsilon. \tag{44}$$

Table 2. General description of the sample sets

Sample set	Problem	Features	Sample size
Boston Housing	regression	14	506
Diabets	regression	20	576
Forest Fires	regression	13	517
Servo	regression	4	167
NBA	classification	12	2235

6. COMPUTATIONAL EXPERIMENT

This experiment was performed to analyse the properties of the sample size estimation methods. The experiment consists of three parts. During the first part, size estimations for all sample sets are obtained, given fixed identical parameters of the methods. During the second part, the dependence of the sufficient sample size on the available sample size is investigated. During the third part, the behaviour of methods depending on the alteration of methods parameters is investigated. Five sample sets described in the Table 2 were used as data. Nine methods in the rows of the Table 3 show sample size estimations for the corresponding data sets.

Table 3. Experiment on sample size estimation for various sample sets

Methods and data sets	Boston Housing	Diabetes	Forest Fires	Servo	NBA
Lagrange Multipliers Test	18	25	44	38	218
Likelihood Ratio Test	17	25	43	18	110
Wald Test	66	51	46	76	200
Cross Validation	178	441	172	120	–
Bootstrap	113	117	86	60	405
APVC	98	167	351	20	–
ACC	228	441	346	65	–
ALC	98	267	516	25	–
Utility Function	148	172	206	105	925

6.1. Sample size estimation for various datasets

This part of computation experiment shows how different methods works on different datasets. The experiment uses next datasets: Boston Housing (Harrison et al. 1978), Diabetes, Forest Fires, Servo (Quinlan 1992), NBA. The result is presented in Table 3. The symbol "–" in the table means that there is not enough data for the prediction.

Each method was provided with the whole sample at the start was performed. The parameters of each method for all samples are registered and described in the Table 4. Since the Lagrange, Likelihood Ration and Wald tests are asymptotic equivalent the parameters of these methods were set identically. The parameters of the Average Coverage and Average Length methods were set identically as well.

Table 4. List of parameters of the sample size estimation methods

Method	GLM parameters	l	ε	α	β
Lagrange Multipliers Test	$\mathbf{w}_u^0$	–	0.2	0.05	0.2
Likelihood Ratio Test	$\mathbf{w}_u^0$	–	0.2	0.05	0.2
Wald Test	$\mathbf{w}_u^0$	–	0.2	0.05	0.2
Cross Validation	–	–	0.05	–	–
Bootstrap	–	0.5	–	0.05	–
APVC	–	0.5	–	–	–
ACC	–	0.25	–	0.05	–
ALC	–	0.5	–	0.05	–
Utility function	–	–	0.005	–	–

6.2. Dependency between available data and sample size prediction

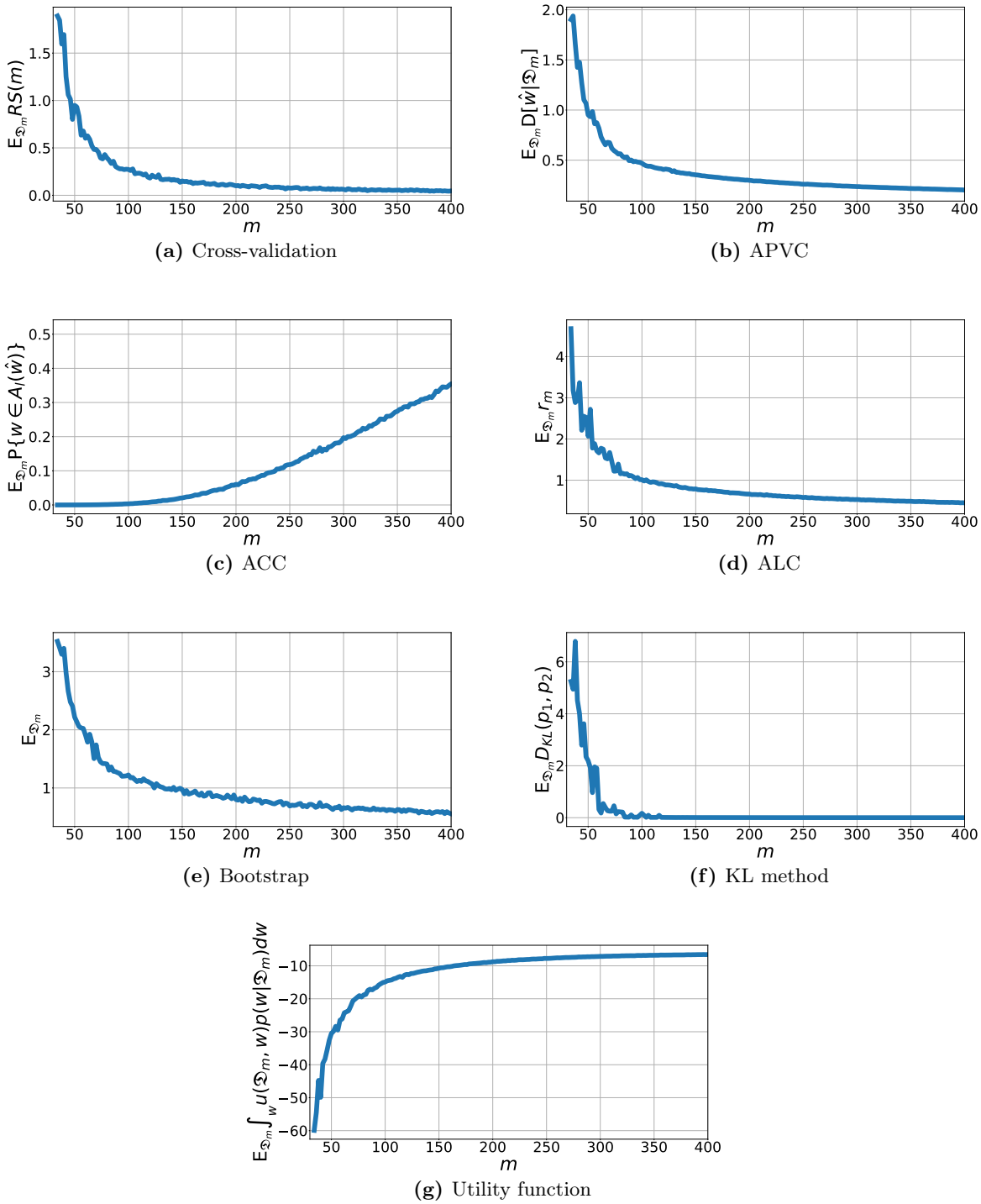
The computational experiment was conducted to analyse the described methods. The sample set is the Boston Housing Dataset. Having a full sample set, fix some sample size m and generate series of bootstrap subsamples of size m from the initial sample set. For different values of m compute m^*, average them and calculate standard deviation.

The Figure 1 shows the dependence of the static values of each method for a given dataset with a fixed sample size m. The thresholds for each method are set expertly, which allows us to control different features of the dataset. The figure 1 demonstrates the adequacy of different definitions of sample size sufficiency. All the presented function are monotonous and all of them are asymptotically tend to a constant. The Figure 2 shows methods' results on samples of different size. It shows how methods differ in variance of computed m^* and behaviour in case of small sample set. The methods converge and the result become independent of sample size from some value of m.

Small variance interpret as stability of the methods output with little dependency on a particular subsample of some size. Some of the methods can not give estimation of sufficient sample size if they don't have such sample. That means that they are useless in terms of prediction, but can be used for retrospection and analysis of already conducted experiment.

6.3. Experiment on subsamples

The dependence between the sample size estimation with the help of a certain method and the volume of data available to this method were considered in this exepriment. The constant achieved by the dependence diagram m^* on m is the forecast of the optimal sample size method. If this constant is less than m where the diagram achieved it, then the method forecasts the optimal sample size before obtaining it. Only Lagrange, Wald methods, and the likelihood ratio method have such property.



(a) Cross-validation **(b)** APVC **(c)** ACC **(d)** ALC **(e)** Bootstrap **(f)** KL method **(g)** Utility function

Figure 1. Methods main scalar functions behaviour dependent on available sample size

6.4. Experiment on hyperparameter alteration

The alteration of the sample size estimation depending on the alteration of certain hyperparameters for Bayesian methods, as well as the methods based on cross-validation and bootstrap is investigated in this experiment. In order to analyse the methods behaviour, see the sample Boston Housing, the other samples have the identical tendency.

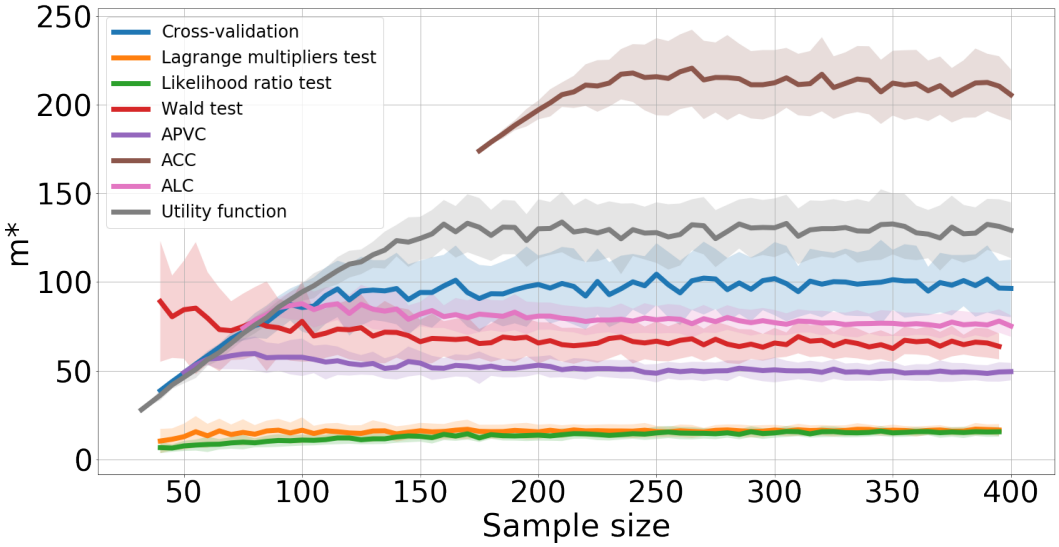


Figure 2. Methods behaviour depending on the available sample size

Bayesian methods, as well as the methods based on cross-validation and bootstrap work on the basis of a certain decision rule for a certain sample function. In the Figure 1, the dependence of these functions on the subsample size is shown. As shown in the Figure 1, these functions are monotonically decreasing, or increasing. The function type of behaviour depends on the method. By altering the restrictions set by the application task, the sample size which will comply with these restrictions can be altered.

7. SOFTWARE

Methods of sample size estimation listed above are implemented in a simple Python package. This package can be used both for prediction sufficient sample size in the early stage of the experiment and for retrospective analysis of the sufficiency of the sample set. This package also includes some auxiliary functions for sample size research and visualization of the results such as on the figures above. Source code for sample size estimation service are available at `https://github.com/andriygav/SampleSizeLib`. Experiment code and datasets used in the paper are available at `https://github.com/ttgadaev/SampleSizeEstimation`.

8. CONCLUSION

In this work, the methods for determining the optimal sample size with the help of frequency and Bayesian methods were investigated. A simulation experiment was performed to investigate the properties of these methods.

As shown in the experiment, the methods used have a series of restrictions. Frequency methods based on the Lagrange tests, likelihood ratio, and Wald tests are asymptotic and cannot be used with little quantity of objects. Bayesian methods, as well as the methods based on cross-validation and bootstrap analyse a certain sample function which requires the greater quantity of objects in order to determine the sample size.

All methods considered require the greater quantity of objects in order to determine the sufficient sample size. It is desired to obtain a criterion which forecasts certain estimation of the sufficient sample size for a smaller quantity of objects, which would make it possible to determine the required quantity of objects for experiment at the early stages of data collection.

9. FUTURE WORK

The main problem is to find a sample size estimation method, that is both interpretable and able to give estimations higher than the available sample set. One of the approaches is to combine Bayesian approach with estimation of statistic on the insufficient sample set. Vector of model parameters are distributed normally. Our future work is to construct a statistic, which estimates mean and covariance matrix of parameters distribution. Such statistic should be estimated on an insufficient sample size in terms of one of the Bayesian definitions of sufficiency.

Acknowledgments. This research was supported by Russian Foundation for Basic Research (projects 19-07-01155, 19-07-00875) and National Technological Initiative (project 13/1251/2018).

REFERENCES

1. Demidenko E (2007) Sample size determination for logistic regression revisited. *Statist. Med.* 26:3385–3397.
2. Harrison D, Rubinfeld D (1978) Hedonic prices and the demand for clean air. *Economics and Management* 5:81–102.
3. Joseph L, Berger R, Be'lisle P (1995) Bayesian and mixed bayesian likelihood criteria for sample size determination. *Statistician* 16:769–781.
4. Joseph L, Wolfson D, Berger R (1997) Sample size calculations for binomial proportions via highest posterior density intervals. *Statistical Medicine* 44:143–154.
5. Kloek T (1975). Note on a large-sample result in specification analysis. *Econometrica* 43:933–936.
6. Lindley D (1997) The choice of sample size. *The Statistician* 46:129–138.
7. Motrenko A, Strijov V, Weber G (2014) Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics* 255:743–752.
8. Quinlan J (1992) Learning with continuous classes. *Proc. 5th Australian Joint Conference on AI* 343–348.
9. Qumsiyeh M (2013) Using the bootstrap for estimation the sample size in statistical experiments. *Journal of modern applied statistical methods* 8:305–321.
10. Rubin D, Stern H (1998) Sample size determination using posterior predictive distributions. *Sankhya: The Indian Journal of Statistics Special Issue on Bayesian Analysis* 60:161–175.
11. Self S, Mauritsen R (1988) Power/sample size calculations for generalized linear models. *Biometrics* 44:79–86.
12. Self S, Mauritsen R, Ohara J (1992) Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* 48:31–39.
13. Shieh G (2000) On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* 56:1192–1196.
14. Shieh G (2005) On power and sample size calculations for Wald tests in generalized linear models. *Journal of Statistical Planning and Inference* 128:43–59.
15. Wang F, Gelfand A (2002) A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. *Statistical Science* 17:193–208.