DIALOG

Distributed Intercept Adjusted LOGistic regression

Hrytsenia, M.

Award date:
2019

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science

DIALOG - Distributed Intercept Adjusted LOGistic regression

Master Thesis

Maksim Hrytsenia

Supervisors:
TU/E: full pr. dr. Edwin van den Heuvel
IKNL: dr. Gijs Gelejnse
TU/E: dr. Zhuozhao Zhan

Eindhoven, August, 2019

Abstract

For long time due to the latest development in the medical domain people's length of life in Europe continues to increase. Because of these developments elderly people every year get new additional days to live. Unfortunately, this tendency also increases the frequency of some serious diseases that elderly people tend to have, for example, cancer. For that reason, importance of studies in the cancer domain increases as well. Different outcomes of the cancer can be predicted in advance, for example, surgery or chemotherapy. These predictions are providing valuable insights for doctors and therefore improve life of ill patients. However, for making predictions there is a need for a lot of data, which is, unfortunately, not the case for a number of countries in Europe with small population, for instance, Iceland, Luxembourg and Slovenia.

These countries need to participate in multinational studies in order to help their people. Additionally, even if a country has a relatively high population for example, Netherlands, for some rare events it also requires a multi-country cooperation to get enough data and as a result reliable estimations or predictions. Since the data of patients is protected by number of laws, for example, GDPR, this analysis can not be simply done in one location. One of the main ways to perform this analysis is by using distributed(federated) privacy-preserving algorithms. This algorithms on the top of proper secure infrastructure are able to provide reliable results by running at separate machines which are located in different countries sharing small not-private pieces of information.

Chemotherapy or surgery are binary outcome which can be well captured by logistic regression. Luckily, there is a number of different distributed logistic regression methods described in the literature. However, the problem is currently they are not evaluated under the same scenarios and from different perspectives together. Therefore it is not obvious which method is the most appropriate for each particular situation.

Furthermore, due to the population dissimilarities(such as average length of life), different countries would have specific patterns which are common only for them, assessing which requires country specific intercepts. Previously, performance of logistic regression algorithms described in the literature was not evaluated for this particular setting.

For that reason, this thesis has two main targets: 1)evaluate performance of existing distributed logistic regression algorithms and 2)find out if current models are able to calculate site specific intercepts.

Acknowledgments

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

With more patient's clinical information available in electronic format healthcare researches started to investigate the influence of different factors on patient's mortality and disease. Sometimes data possessed by one hospital or research center is enough to perform a valid statistical analysis. However, in most of the cases, this data is not enough to guarantee valid statistical results.

For that reason, for long time scientists from research centers and hospitals have used different techniques which allowed to combine data together. These joint datasets allowed to get valid statistical results. However, due to the recent privacy concerns and enforced regulations, conducting a large-scale collaborative research started to be almost impossible. Therefore, there is an increasing need in methods which would allow to perform a collaborative study without sharing private data.

Distributed(Federated) learning is a promising alternative to perform multi-institutional analysis without exchanging personal data. Instead of personal data, algorithms share only summary statistics which can not be tracked back to individual patients. Distributed learning allows development of privacy-preserving algorithms with high accuracy. These algorithms can be developed to perform different tasks such as linear and logistic regression [7].

One of the main fields in the medical domain, where distributed learning may solve problems and save lives, is a cancer research. Cancer has many different types with varying occurrences. For example, according to the European Cancer Information System (ECIS) Nasopharynx type of cancer has had approximately 100 number of incidents in the Netherlands in 2018 [1]. This number of records does not allow to achieve reliable results using only data from the Netherlands. For that cancer type distributed learning algorithm is a promising solution to improve the cancer care by enabling cooperation with other countries.

One of the main tasks that researchers from cancer domain carry out is the estimation of patients survival rate after some time period. Survival outcome of patient in that case is captured by binary variable. Therefore, this problem models a distribution of binary variable and one of the most known and suitable algorithms for this task is a logistic regression model. Simple interpretation of this statistical model allows doctors to easily estimate importance and influence of different factors.

As was mentioned previously, for the rare types of cancer international cooperation is required to perform the valid statistical analysis. However, comparing different countries in Europe it is easy to see that they have differences in many aspects, for example, average length of life. In that case, inclusion and estimation of this country variability is a key aspect to have a valid statistical model. It is the main task of a DIALOG project.

Figure 1.1: Male kidney cancer survivability in France, Germany and the Netherlands

1.2 Problem statement

Previously researchers have concentrated on investigating the distributed logistic model for the data assuming that model is the same for different location. However, this assumption does not always hold.

To provide the evidence we have used the information from ECIS. For example, let consider the kidney cancer survivability in 3 almost equally developed Western European countries: France, Germany and the Netherlands [2]. The survivability for different age groups is shown in the Figure 1.1. This plot shows that these countries have almost consistent survivability country heterogeneity for all age groups. This heterogeneity may point out that there is a consistent difference between the countries which is dependent on some country factors and not on the cancer treatment. These possible factors can include, for instance, different live length or eating habits.

These differences prove the importance of their inclusion into the distributed models. However, to the current date, there has been no investigation of these type of model in the literature. Therefore, there is a need to find a way how to incorporate these differences into the model.

1.3 Aims and scope

The main aim of this study is to find out if existing distributed logistic regression models can estimate the underlying differences between different locations under different scenarios. This research incorporates this difference by using a site-specific intercept which has a distinct value for each of the location. Particularly, we are interested in the underlying collaboration network which includes from 2 to 9 different research centers. The scope is limited to the horizontal data partitioning which allows to use records from different sites to achieve better statistical results.

This research includes distributed logistic regression models which do not exchange patient data in any format. This methods scope allows to concentrate on the solutions which can be run under almost any privacy regulations.

1.4 Research Questions and our Hypothesis

Considering the aims and scope of this research, this thesis is dedicated to answer the following research questions:

1. What are the existing distributed logistic regression models described in the literature? Do these techniques preserve the privacy of individual data records?
2. How well existing distributed logistic regression models fit the data from different sites when some of sites do not have enough data to construct a reliable model on their own?
3. How well site-specific intercept is assessed by existing techniques and how these techniques are compared among each other under different data allocation and generation scenarios?
4. Are the existing techniques able to correctly assess this intercept if binary outcome distribution is not even?
5. Are existing techniques able to create the site-specific intercept if some sites do not have one of the outcomes present on the site?
6. Is the implementation of existing techniques efficient in terms of computational power required?

1.5 Significance

This thesis extends and evaluates of new set of possible data models for the distributed logistic regression models. These data models include site-specific intercept which can capture the underlying data location difference. Previously, these models have not been assessed in the literature.

Additionally, this research provides an up to date classification of distributed logistic regression models. Distributed privacy-preserving logistic regression models recently received a boost from a research community. However, up to the current moment, except [7] no attempt has been made to compare the qualities of proposed models. Therefore, this research extends and updates the methods classification which was made in [7].

1.6 Thesis overview

The rest of the thesis is structured as following: in Chapter 2 required underlying information would be briefly presented to introduce audience to the topic, in Chapter 3 methods classification and selection is discussed, in Chapter 4 modifications of distributed logistic regression models which would allow to capture site-specific intercept are proposed, Chapter 5 includes the main information about simulation which were carried out to estimate the validity of selected models, Chapter **??** presents simulation results, Chapter **??** evaluates the results and **??** concludes the thesis by presenting and discussing the limitations of this research.

Chapter 2

Background

This background chapter is dedicated to make the reader familiar with the overall context of the field and this research particularly. It highlights some of the main difficulties which statistical and computer science researchers in cancer domain have nowadays and briefly demonstrates how the topic of this thesis is linked to that problems. Additionally, this chapter provides theoretical background to the problem described in 1 which can help the reader to understand material which would follow in next sections easier.

To complete these aims, this chapter is structured in the following way: firstly, we would describe a distributed(federated) learning paradigm in the medical domain and its importance, secondly, we would provide some details about the logistic regression model and describe the main features of that techniques, thirdly, we would explain what are the main problems which logistic regression has and how these problems look from data point of view and, lastly, we would conclude this chapter.

2.1 Distributed learning

The main aim of a distributed learning in the medical domain is allowing researchers to construct predicting models based on the data from different locations which can be countries, regions or hospitals. However, the problem to perform computations based on the data from different sources is not completely new and for that reason, for long time researchers tended to use a study-level meta analysis approach(SLMA). Usually, this approach consists of following stages: calculation of summary statistics on each location, transferring data to the central calculation point and evaluation of global model coefficients by using that statistics. However, this approach uses only aggregated information from each of the sources and therefore does not use all the available information and as a consequence is not able to construct models of the same models as based on combined data in single location.

But, in spite of the fact, that combined data can lead to better models, data from different sources can still not be incorporated due to the privacy and ethical reason. Usually, each location(for example, country or hospital) has its own laws and it is extremely difficult to overcome such complexities and pool all the data to the single storage. In order to overcome the difficulties mentioned previously, medical domain researchers started to create distributed learning systems algorithms for that systems.

The first one of that kind of systems was proposed by members of DataSHIELD project [34]. They proposed to use ideas of parallelized analysis and distributed computing for iteratively estimating the coefficients of generalized linear models while keeping data at its original locations. Their schema has a number of features: 1)Data of individuals is not shared, 2)Data nodes do not send message to each other, 3)Central server and data nodes iteratively communicate with each other: based on received summary statistics from data centers, server sends a model update to data nodes. The schema of their project is demonstrated on Figure 2.1.

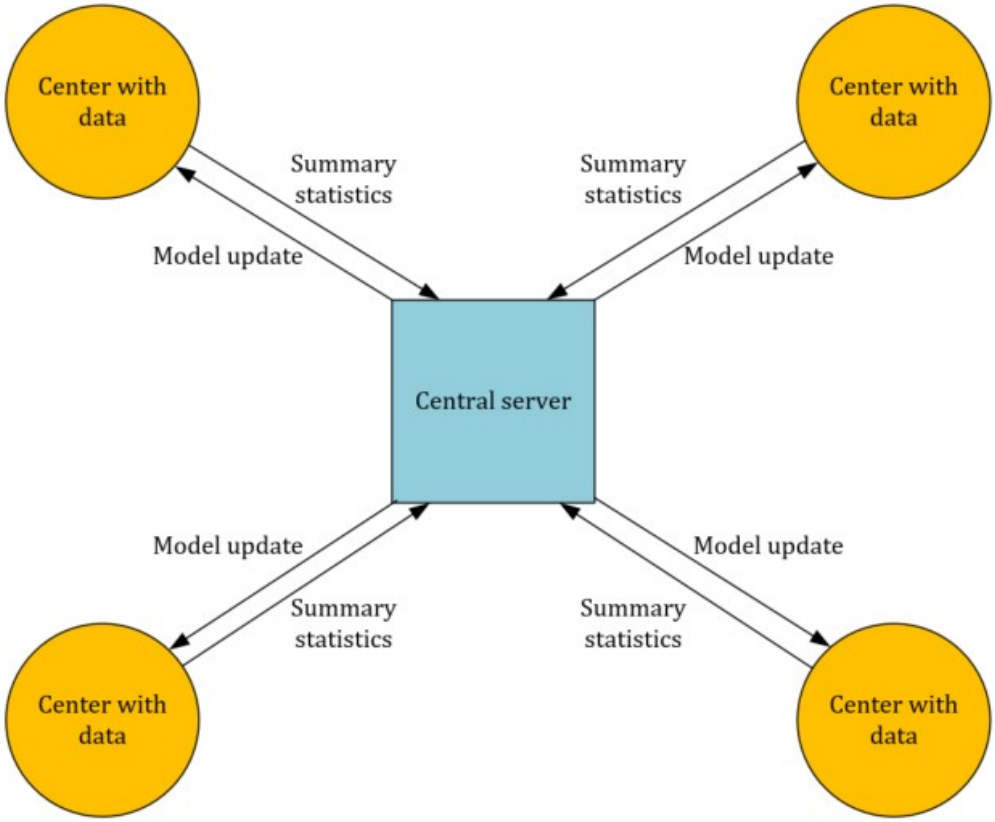


Figure 2.1: DataSHIELD Schema of communication between centers with data and central server

Up to date, due to the recent implementation of GDPR law in the EU and increasing demand for the multi-country research, number of systems to allow distributed learning continues to grow. The examples of this systems include but not limited to euroCAT[8], DataShield [13],[34] and pScanner [28]. Currently these systems are able to calculate many some of popular statistical models and machine learning algorithms and currently continue to propose new ones to further facilitate collaboration between researchers around the globe.

2.2 Logistic regression

Regression models are the essential part of any data analysis when there is a need to explain the relationship between the explanatory(independent) variables and outcome(dependent) variable [15] in the most simple and accurate way. In many situations, the outcome variable may take only binary variables, for example, in the situation if there is a need to find out whether patient has a specific disease or not, the binary variable should be used for modelling. Logistic regression is one of the most used regression models for such type of situations. In the following section we would provide the mathematical background behind this model.

The logistic regression model is using the following formula:

$$p(y = 1|x) = \frac{e^{\alpha+\beta\cdot x}}{1 + e^{\alpha+\beta\cdot x}} \tag{2.1}$$

where $p(y = 1|x)$ means the probability of positive outcome for a data record, *alpha* is an intercept, β is a vector of coefficients which model is estimating and x is a d-dimensional vector of features for each data record. Later in this section we assume that α is included inside β and x_i includes the according intercept feature with all values equal to 1.

The most common method which is used for estimation of coefficients is *maximum likelihood method*. This method try to find the set of values for the coefficients which maximizes the probability to obtain the observed data[15] and this method needs a correspondent likelihood function. This function for logistic regression can be described by there following formula:

$$l(\beta) = \prod_{i=1}^{n} p(y = 1|x_i)^{y_i} \cdot (1 - p(y = 1|x_i))^{1-y_i} \tag{2.2}$$

where y_i - is an outcome value of row i.

To mathematically simplify the calculations, log-likelihood function is usually transformed to the log likelihood:

$$L(\beta) = \ln(l(\beta)) = \sum_{i=1}^{n} (y_i \cdot \ln(p(y = 1|x_i)) + (1 - y_i) \cdot \ln(1 - p(y = 1|x_i))) \tag{2.3}$$

which can be further transformed into

$$L(\beta) = \beta \cdot \sum_{i=1}^{n} x_i y_i - \ln\big(1 + e^{\beta \cdot x_i}\big) \tag{2.4}$$

To find the coefficients which maximize the function we need to differentiate log-likelihood function with respect to β and set the resulting expression to 0. Below we provide the result of differentiation:

$$\frac{\mathrm{d}L(\beta)}{\mathrm{d}\beta} = \sum_{i=1}^{n} x_i y_i - x_i p(y = 1|x_i) = 0 \tag{2.5}$$

Usually the equation below does not have a direct solution and numerical methods need to be used in order to solve that mathematical equation [3]. The most common method which is used is Newton-Raphson algorithm. This algorithm requires the first and second order derivatives with respect to β. Let $S(\beta)$ be the set of first order derivatives and and $I(\beta)$ be the set of second order derivatives:

$$S(\beta) = \frac{\mathrm{d}L(\beta)}{\mathrm{d}\beta} = \sum_{i=1}^{n} x_i y_i - x_i p(y = 1|x_i) \tag{2.6}$$

$$I(\beta) = \frac{\mathrm{d}^2 L(\beta)}{\mathrm{d}\beta^2} = -\sum_{i=1}^{n} x_i x_i' p(y = 1|x_i)(1 - p(y = 1|x_i)) \tag{2.7}$$

Then the iterative update of coefficients is shown on the following formula:

$$\beta_{j+1} = \beta_j - I(\beta)^{-1} S(\beta) \tag{2.8}$$

where j means the number of the iteration.

The algorithm continue to iterate until either 1)the update start to have a very small value which means that algorithm converged or 2)algorithm reaches the maximum number of iterations but still did not converge.

2.3 Complete and quasi-complete separations

According to [3], the most common problems when estimating the logistic regression model coefficients are complete and quasi-complete data separation patterns.

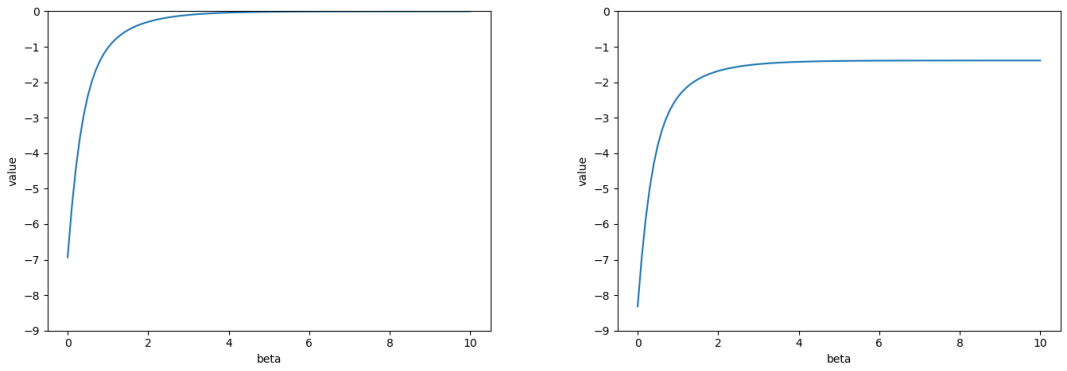
As was described in the previous section, one of the key ideas of logistic regression is maximization of log-likelihood. Complete and quasi-complete data separations are happening when maximum does not exist. It is can be shown using the example from [3]. We would like to demonstrate examples of complete and quasi-complete separations using the set of data from Table 2.3. In this example we can see that variable x is a perfect predictor for the outcome on the left dataset and almost perfect on the right where 2 new rows with $x = 0$ and $y = 0$ and $y = 1$ are added.

x	y
-5	0
-4	0
-3	0
-2	0
-1	0
1	1
2	1
3	1
4	1
5	1

x	y
-5	0
-4	0
-3	0
-2	0
-1	0
0	0
0	1
1	1
2	1
3	1
4	1
5	1

Table 2.3: Example of complete(left) and quasi-complete(right) separations of data

The log-likelihood function of the logistic regression based on these data is shown on Figure 2.2.There we can see that under complete data separation log-likelihood never reaches its maximum value and always stays below 0 so there is no maximum likelihood value. The same holds for log-likelihood under quasi-complete separation but max-value is slightly below than -1.38.



(a) Log-likelihood values under complete separation

(b) Log-likelihood values under quasi-complete separation

Figure 2.2: Log-likelihood function values under complete and quasi-complete separation

In terms of data, complete and quasi-complete data separations are very frequent when either the data set is quite small or when there are extreme divisions in data either in terms of outcome or binary(categorical) variables distribution.

According to [3] there are couple of ways to detect and solve complete and quasi-complete separation problems. What usually is happening when there is a separation in data is that variables with non-existent coefficients start to have very large estimates and standard errors. The solutions for complete and quasi-complete data separations which are suggested include such techniques as 1)combination of categories which does not always work, 2)usage of exact logistic regression which is computationally expensive and 3)usage of Firth maximum likelihood estimation.

2.4 Show how site-specific intercept effect the data on the chart(two lines, one of which is on the top of another)

Chapter 3

Literature review

Logistic regression model is used in different domains for decades. Distributed logistic regression is a new step of evolution which allows closer collaboration and increases the amount of data which can be processed.

At the current moment, to our knowledge, there is no distributed logistic regression model with site-specific intercept. In order to evaluate if current more simple logistic regression models can accurately estimate models with site-specific intercept we have made literature review.

Usually distributed logistic regression is just one of the models which can be programmed using optimization techniques which researchers develop. In overall, there techniques are concentrating only on some special applications. These application are mostly different in terms of priorities which researchers are trying to address. In overall, we found out that these properties can be categorized into 4 main groups: 1)Amount of data which model is able to handle, 2) Running time of the model, 3) Accuracy of the model and 4) Privacy of the data. In the following sections we would provide some overview how different domains are linked to these priorities and what are the priorities which would allow to address the research question from this thesis.

3.1 Literature search

On 10.7.2019 last iteration was done in Google scholar database using the following key words: "decentralized logistic regression", "distributed logistic regression", "multi-party logistic regression" and "federated logistic regression". Relevant papers were included based on their headers and abstracts.

Additionally, the relevant search was also done through the referenced by these papers works. Furthermore, additional search in google scholar was done to select the relevant papers which cited earlier selected works.

3.2 Applications for big data

This family of optimization techniques is concentrated on processing high number of records with usually large number of covariates. These techniques allow to process this data as fast as possible. Therefore, they mostly do not consider the accuracy of the results and only a bit restricted in terms of privacy.

Three of the techniques which received a close attention in the recent time is Alternating Direction Method of Multipliers(ADMM)[5], stochastic coordinate descent(SCD) [20] and stochastic gradient descent(SGD) [6]. These models are concentrating on scenario with multi-dimensional data with many records which is distributed across the network with many agents. Therefore, usually these methods are extremely slow when converged to the high accuracy. At the same time, the main scenarios where this approach is used does not need the high accuracy but moderate is usually sufficient.

The ADMM also boosted a number of applications which further continue to provide more effective solutions for different big data domains where there are a lot of connected agents with small amount of information [31], [25], [29], [26], [33], [23] for instance, social networks and network of mobile devices.

Due to the low accuracy and specific underlying machines network requirements, this family of techniques was not included in DIALOG.

draw a picture of these approaches -¿ social graph + communication

3.3 Applications for cloud computing

This family of techniques is trying to privately collect the information from different devices and run logistic regression model in a single storage. This scenario is extremely common for cloud computing.

These approaches have a number of unique features. First of all, they usually put the information from many databases into one centralized. Secondly, for combining the data in the private way, they encrypt all the data records. Lastly, they use their algorithms on the top of encrypted data. This data and its encryption specify the key property of this approach: achieve high privacy for all individual data records. To achieve this property, homomorphic encryption technique has been widely used [37], [22], [4], [21], [16]. However, there is a pay off in terms of high running time and/or lower accuracy. For these reasons, these models were not included in DIALOG.

draw a picture of these approaches -¿ how data is uploaded and shared

3.4 Approaches for advanced individual and institutional privacy

Techniques which were demonstrated above mostly concentrate on the privacy of individual data records. However, there is one more side of the the privacy concern: privacy of intermediate data and institutional performance.

To address this issue, the conservative approaches were introduced which are able to protect intermediate results and institutional privacy [14], [9]. These techniques demonstrate methods with high privacy guarantees however with inefficient implementation and low accuracy accordingly. Furthermore, they only consider high-sparsity data with many data rows. Due to the inefficiency of that methods there were not included in DIALOG.

3.5 DIALOG methods

As was described in the introduction, focus of DIALOG is small researchers networks where number of centers varies from 2 to 10. Additionally, DIALOG is also concentrate to facilitate the research where individual data centers do not have enough data to construct models on their own. Therefore, DIALOG models include models which can accurately estimate coefficients.

Based on the research selection, there are only 2 models which satisfy the properties above:model which uses Iteratively Re-weighted Least Squares algorithm(Newton-Raphson method) or Communication efficient statistical inference. Both of them updated coefficients iteratively and require a central server to coordinate these iterations. Details of these approaches are provided below.

3.5.1 Distributed logistic regression using Iteratively Re-weighted Least Squares method

One of the first distributed privacy-preserving logistic regression models was the one proposed in [12] and later expanded in [34] and [18] by members of a DataSHIELD project. The algorithm proposed in that papers divides computations into 2 parts: calculation of the summary statistics

on each of the nodes and estimation of the model coefficients on the central server. Data nodes computations are performed based on the local data on the node and produce the summary statistics which is later shared with the central server. At the start of each iteration each of data nodes receives relevant estimated coefficients. These coefficients are supplied by the central server and used for computations. After summary statistics from all sites is received, central server updates coefficients using Iteratively Re-weighted Least Squares (IRLS) and Newton-Raphson algorithms. The central server shares the coefficients with data nodes and by that starts a new iteration. The communication schema between the central site and corresponding data nodes is presented on Figure .. . Below we specify the key properties of that algorithm.

add communication schema

Description

One of the key properties of distributed IRLS and Newton-Raphson method is that they allow to parallelize their computations across different data nodes. By this parallelization computations can be efficiently split across the data network. The application of IRLS and Newton-Raphson algorithms for the distributed logistic regression allows parallelization of computations across the network of data nodes. As was described in Section 2.2, mathematically iterations of Newton-Raphson using centralized logistic regression are explained by the following formula.

$$\beta_{j+1} = \beta_j - I(\beta)^{-1} S(\beta) \tag{3.1}$$

where $I(\beta)$ is an information matrix and $S(\beta)$ is a score vector. The computations of $I(\beta)$ and $S(\beta)$ can be divided across the sites:

$$I(\beta) = \sum_{s=1}^{k} I_s(\beta) \tag{3.2}$$

$$S(\beta) = \sum_{s=1}^{k} S_s(\beta) \tag{3.3}$$

where s is the data site and k is overall number of computational nodes.

Number of operations which central server performs is reduces since computations of summary statistics is distributed across the data nodes. Central server only needs to sum $S(\beta)$ and $I(\beta)$ which are received from different data sites and the inversion of $I(\beta)$ to finish the iteration.

Some readers may point out that inversion of matrix $I(\beta)$ would be the most expensive computationally. This computations complexity is depending on the number of parameters that logistic regression model has. However, as was mentioned in the Introduction, DIALOG is going to estimated models based on the data with low number of covariates and therefore number of independent variables should not significantly effect the computation time.

Properties

One of the key advantages of this algorithm is that it guarantees the similarity of coefficients comparing to the pooled logistic regression as was proved in [18]. Furthermore, as mentioned in [35] number of iterations required to converge is usually small which means that less data is shared between the data node and central server.

However, this approach had also a number of limitations: 1)it requires a permanent communication and coordination between node and the central server which can be inappropriate for some situation from time and security points of consideration, 2)it does not protect the institutional privacy and 3)model needs a retraining when node data is updated or new node is added. Fortunately, these drawbacks are not significant for DIALOG setting.

Privacy guarantees

Since each data node separately performs its calculations and shares only the summary statistics with the central server, the raw data usually can not be recovered and kept in a secure way on each of the data site.

However, nevertheless only summary statistics is shared, SPARK work [11] revealed that this statistics in some scenarios can reveal a sensitive patient's and center's information. Therefore they developed a SPARK protocol which added additive homomorphic encryption system to this distributed logistic regression algorithm. By adding homomorphic encryption it allowed to protect privacy both of centers and enhanced privacy of individuals. Similar adjustment to enhance privacy was also proposed in [30].

One of the alternative models how to protect privacy of the centers is to apply secure summation algorithm as was proposed in [35]. To preserve the anonymity of data centers, for each iteration it uses a random matrix which hires the intermediate results of each individual data center.

In DIALOG, these adjustments are not implemented since we assume that they are part of the underlying infrastructure rather than the algorithm itself. The most important fact is the existence of techniques. In this case, if additional privacy protection mechanism is required, it can be easily implemented in the distributed learning infrastructure.

Improvements

There are still some problems in the distributed logistic regression that are not addressed. For example, it is difficult to measure the performance and accuracy of coefficients which distributed model provides. Unfortunately, not so much work has been done in this direction.

Currently there is still only two distributed techniques proposed in the literature, which would allow to measure the model performance in a distributed way. In [12] it was proposed to compare fit of the models by using distributed calculation of log-likelihood and Pearson χ^2. Also authors of GLORE [35] have demonstrated how to calculate an H-L test and the AUC in a distributed privacy-preserving way. These methods are essential for estimating goodness-of-fit and discrimination performance of predictive models in a distributed privacy preserving way.

To handle clustered and correlated data, Generalized Linear Mixed Model was proposed in [17]. However, we assume that in our data scenario we would not have correlated data. For that reason, this method was not included in DIALOG.

Regularized logistic regression is an extension of logistic regression which allows better generalisation of logistic regression model. There are a couple of proposed privacy-preserving regularized algorithms which can extend distributed logistic regression model by using $l2$ regularization [36], [24]. Coefficients of these models are also updated using distributed IRLS algorithm. For that reason, when regularized term is removed from these models, they become equivalent to the distributed IRLS method described earlier. Since DIALOG model set up does not include penalized terms, we have removed these models from further investigation.

3.5.2 Communication-Efficient distributed statistical inference

One of the approaches that can further reduce the information which is shared between data sites and central server is Communication-Efficient distributed statistical inference(CSL). Originally, the framework was introduced by [32] but performance of distributed logistic regression based on this framework was later evaluated by [19]. This approach requires to transfer less information and under some conditions should get the similar accuracy comparing to the distributed IRLS method.

Description

Logistic regression algorithms are deriving its coefficients by minimizing its global loss function. This algorithms is using an idea of Taylor expansion of the global loss function. Using that method it is able to minimize the amount of summary statistics required. Distributed IRLS requires the

computation of Score vector(first-order derivative of log-likelihood function) and Hessian(second-order derivative) at data sites as CSL requires computation of first-order derivatives only. Additionally, it also uses the data from one of the data nodes to construct global loss function(without loss of generality, let assume it is the site 1). It is different from the distributed IRLS method where central server did not need to have its own data.

Properties

The main advantages of CSL are 1)light computations on the data nodes comparing to distributed IRLS method from Section 3.5.1 since only calculation of first-order derivative is needed, 2)small amount of information which is shared across the network and 3)small number of iterations required to converge in the most of the cases. Furthermore, under some scenarios, only one iteration to calculate global gradient and estimate global loss function is sufficient to achieve high accuracy as was mentioned in One-shot distributed algorithm to perform logistic regression(ODAL) [10]. Taking into account the possible advantages of ODAL, DIALOG also includes implementation of it.

However, CSL also has disadvantages. To successfully converge to the reliable estimates, number of machines k should be lower than $\sqrt{N}$ where N is total number of records. Additional restriction to achieve high accuracy specifies that only when number of records on the site 1 is much higher than number of data covatiates, CSL is achieving the optimal statistical accuracy. Fortunately, due to the application which was specified for DIALOG in 1, total number of data nodes is not high and this records distribution should always satisfy these requirements.

Privacy guarantees

Since CSL shares only first-order derivatives of size d, where d is number of data covariates, across the network, it protects the local individual data from identification by other sites. Additionally, it also requires a small number of iterations to converge in most of the cases, which makes it almost impossible to deliver the information about individual records based on so small amount of information.

3.6 Comparison of evaluations

The main properties of selected methods can be seen in the Table 3.1.

Method	Data	Accuracy	Running time	Shared data	Privacy preservation
Distributed IRLS	Can be very time and space demanding if number of covariates is high	Produces accurate estimates under almost all scenarios	Can be extemely slow if number of covariates is high	First and second order log-likelihood derivatives	Individual information can be revealed under rear scenarios
CSL	Can efficiently handle big dataset with high number of covariates	Produces accurate estimates when: 1) central site has enough records 2) number of covariates is smaller than total number of records	Fast	First order derivative of log-likelihood	Individual information is protected
ODAL	Can efficiently handle big dataset with high number of covariates	Produces accurate estimates when: 1) central site has enough records 2) number of covariates is smaller than total number of records	Very fast	First order derivative of log-likelihood	Individual information is protected

Table 3.1: Comparison of selected methods for DIALOG

Chapter 4

Methods

4.1 Site specific intercept. Data view

DIALOG is intended to estimate coefficients of distributed logistic regression model which includes site-specific intercept. As was mentioned earlier in Chapter 3, currently there is no models of that kind which were researched earlier. Therefore DIALOG includes distributed logistic algorithms which estimate more simple model. The outcome of this model can be represented as $y_{si} \in \{0, 1\}$ where y_{si} represents the outcome for subject i which is located on site s. The outcome for this subject is estimated by the model as

$$E(y_{si}) = p(y = 1|x_{si}) = \frac{e^{\alpha + x_{si}^T \beta}}{1 + e^{\alpha + x_{si}^T \beta}} \tag{4.1}$$

where x_{si} specifies the covariates vector of record i located on site s, β represents coefficients for that covariates and α is an intercept which is shared among all sites.

To capture the site-specific intercept, model needs to be more complex. Site-specific intercept can be included in that model by adding intercept α_s which is specific for each site. Then the model from 4.1 can be rewritten as:

$$E(y_{si}) = p(y = 1|x_{si}) = \frac{e^{\alpha + \alpha_s + x_{si}^T \beta}}{1 + e^{\alpha + \alpha_s + x_{si}^T \beta}} \tag{4.2}$$

or in the logit form:

$$\log\left(\frac{p(y = 1|x_{si})}{1 - p(y = 1|x_{si})}\right) = \alpha + \alpha_s + \sum_{k=1}^{p} \beta_k \cdot x_{ijk} \tag{4.3}$$

where α_s is a specific intercept for site s, X_{ijk} is a value of covariate k for subject j on site i and β_k is a covariate k parameter which is shared for all subjects on all sites.

Let consider an example. Suppose that there is a dataset with only one covariate Age which is distributed across 2 sites. Assume that each of the sites has 50 records, then this data can be partitioned as demonstrated in Table 4.1.

Row	Intercept	Age
1	1	$a_{1,1}$
2	1	$a_{1,2}$
...	...	...
50	1	$a_{1,50}$

Row	Intercept	Age
1	1	$a_{2,1}$
2	1	$a_{2,2}$
...	...	...
50	1	$a_{2,50}$

Table 4.1: Data without site-specific intercepts on sites 1 and 2 accordingly

In order to add a specific intercept for each of the sites there is a need to add columns C_s $|s \in (2, k)$ to all records in the dataset, where s means specific site and k - overall number of sites.

Each of these columns identifies a site to which a data record belongs. There is no need to add an additional intercept for site 1 since the differences between sites are already captured by specific intercepts added for other sites.

Therefore, for the demonstrated example with 2 sites, there is a need to add one extra-column C_2 to the data on each of the sites and fill it with value 1 on site 2. After the column addition the dataset takes the form as presented in Table 4.2.

Row	**Intercept**	**Age**	**C_2**
1	1	$a_{1,1}$	0
2	1	$a_{1,2}$	0
...	...	...	...
50	1	$a_{1,50}$	0

Row	**Intercept**	**Age**	**C_2**
1	1	$a_{2,1}$	1
2	1	$a_{2,2}$	1
...	...	...	...
50	1	$a_{2,50}$	1

Table 4.2: Data with site-specific intercepts on sites 1 and 2 accordingly

In this case, the logit for each data record on sites 1 and 2 can be rewritten as:

$$\log\left(\frac{p(y=1|x_{1i})}{1-p(y=1|x_{1i})}\right) = \alpha + \beta_{Age} \cdot a_{1,i} + \beta_{C_2} * 0 = \alpha_1 + \beta_{Age} \cdot a_{1,i} \tag{4.4}$$

and

$$\log\left(\frac{p(y=1|x_{2i})}{1-p(y=1|x_{2i})}\right) = \alpha+\beta_{Age}\cdot a_{2,i}+\beta_{C_2}*1 = (\alpha+\beta_{C_2}*1)+\beta_{Age}\cdot a_{2,i} = \alpha_2+\beta_{Age}\cdot a_{2,i} \tag{4.5}$$

where $i \in \{1, \ldots, 50\}$.

Later in this thesis specific and shared intercepts are combined together for simplicity purposes and would be identified as α_s where s means a specific site.

4.2 Distributed IRLS for model with site-specific intercept

The main computations which centralized IRLS algorithm performs are calculations of $S(\beta)$ and $I(\beta)$. Their values can be calculated according to equations 4.6 and 4.7.

$$S(\beta) = \sum_{i=1}^{n} x_i y_i - x_i p(y=1|x_i) \tag{4.6}$$

$$I(\beta) = -\sum_{i=1}^{n} x_i x_i^T p(y=1|x_i)(1-p(y=1|x_i)) \tag{4.7}$$

Distributed IRLS as was mentioned in section 3.5.1, split computations of $S(\beta)$ and $I(\beta)$ so they can be done locally on each of the sites with corresponding data. Let assume that there are k sites with data and each of the sites has n_s data records where s is a site. Then equations 4.6 and 4.7 can be rewritten as:

$$S(\alpha,\beta) = \sum_{s=1}^{k}\sum_{i=1}^{n_s} x_{si} y_{si} - x_{si} p(y=1|x_{si}) = \sum_{s=1}^{k}\sum_{i=1}^{n_s} x_{si} y_{si} - x_{si}\frac{e^{\alpha_s + x_{si}^T\beta}}{1+e^{\alpha_s+x_{si}^T\beta}} \tag{4.8}$$

$$I(\alpha,\beta) = -\sum_{s=1}^{k}\sum_{i=1}^{n_s} x_{si}x_{si}^T p(y=1|x_{si})(1-p(y=1|x_{si})) = $$
$$= -\sum_{s=1}^{k}\sum_{i=1}^{n_s} x_{si}x_{si}^T\frac{e^{\alpha_s+x_{si}^T\beta}}{1+e^{\alpha_s+x_{si}^T\beta}}(1-\frac{e^{\alpha_s+x_{si}^T\beta}}{1+e^{\alpha_s+x_{si}^T\beta}}) \tag{4.9}$$

4.2.1 Initialization

Coefficients β and α converge to the final value taking into account previous value as it is shown in equation 3.1. Therefore, initial coefficients need to be assigned before the start of iterations. There are many possible strategies but usually setting all coefficients to 0 is a good choice in the most of the cases.

4.2.2 Algorithm

In overall, the calculations which data sites and central server perform during completion of IRLS method can be summarized by the following algorithm:

$\alpha_0 = 0, \beta_0 = 0$
$last =$ maximum number of iterations
for $j = 1$ **to** $last$ **do**
  Send $\alpha_{j-1}, \beta_{j-1}$ to all data nodes
  for $s = 1$ **to** k **do**
    $I_s(\alpha_{j-1}, \beta_{j-1}) = \sum_{i=1}^{n_s} x_{si} x_{si}^T \frac{e^{\alpha_s + x_{si}^T \beta}}{1+e^{\alpha_s + x_{si}^T \beta}} (1 - p(y=1|x_{si}) p(y=1|x_{si}))$;
    Send $I_s(\alpha_{j-1}, \beta_{j-1})$ and $S_s(\alpha_{j-1}, \beta_{j-1})$ to the central server;
  end
  $I(\alpha_{j-1}, \beta_{j-1}) = \sum_{s=1}^{k} I_s(\alpha_{j-1}, \beta_{j-1})$
  $S(\alpha_{j-1}, \beta_{j-1}) = \sum_{s=1}^{k} S_s(\alpha_{j-1}, \beta_{j-1})$
  $\alpha_j \oplus \beta_j = (\alpha_{j-1} \oplus \beta_{j-1}) - I(\alpha_{j-1}, \beta j - 1)^{-1} S(\alpha_{j-1}, \beta_{j-1})$
  if *algorithm converged* **then**
    $last = j$;
  end
end

Algorithm 1: Distributed IRLS algorithm

where $\oplus$ means vectors concatenation.

The process continues until coefficient values converge or maximum number of iterations is reached. Convergence criteria and maximum number iterations need to be specified for each specific scenario where algorithm is going to be used.

4.3 CSL and ODAL for model with site-specific intercept

In this section we would explain the CSL algorithm in more detail. First for clarity purpose we would start with a model without a site-specific intercept which would be later excluded.

The main point of CSL algorithm is simliar to IRLS: minimization of the cost function 4.10.

$$\widetilde{\beta} = \underset{\beta}{\operatorname{argmin}} E[L(\beta)] \tag{4.10}$$

where $\widetilde{\beta}$ is a set of covariates coefficients which model is estimating and $L(\beta)$ is global loss function which is calculated as

$$L(\beta) = \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + e^{x_i^T \beta}\right) - y_i x_i^T \beta \tag{4.11}$$

where N is total number of records on all sites, x_i is a covariates vector of a record i and y_i is an outcome of a record i.

After that the global loss function is expanded into infinite series:

$$L(\beta) = L(\overline{\beta}) + \nabla L(\overline{\beta})(\beta - \overline{\beta}) + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j L(\overline{\beta})(\beta - \overline{\beta})^{\otimes j} \tag{4.12}$$

where $\overline{\beta}$ is any initial estimate of β.

This series later is approximated using local derivatives instead of global derivatives leading to the following *surrogate likelihood function*:

$$\widetilde{L}(\beta) = L(\overline{\beta}) + \nabla L(\overline{\beta})(\beta - \overline{\beta}) + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j L_1(\overline{\beta})(\beta - \overline{\beta})^{\otimes j} \tag{4.13}$$

Then by using a Taylor expansion of $L_1(\beta)$ around β and omitting the additive constants the surrogate likelihood takes the following form:

$$\widetilde{L}(\beta) = L_1(\beta) - (\nabla L_1(\overline{\beta}) - \nabla L(\overline{\beta}))\beta \tag{4.14}$$

where $L_1(\beta)$ - local loss function calculated at the first machine according to the equation 4.11, $\nabla L(\overline{\beta})$ is a global gradient calculated as $\nabla L(\overline{\beta}) = \sum_{s=1}^{k} \nabla L_s(\overline{\beta})$ where s means the site and k - overall number of sites. Gradient at each site is calculated according to the equation 4.15.

$$\nabla L_s(\overline{\beta}) = \frac{1}{n_s} \sum_{n=1}^{n_s} (\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} - y_n) x_n \tag{4.15}$$

where n_s means number of records which site s has.

Then applying this formulas for a model with a site specific intercept, local loss function at site 1 and gradients can be calculated according to equations 4.16 and 4.17.

$$L_1(\alpha, \beta) = \frac{1}{n_1} \sum_{i=1}^{n_1} \log\left(1 + e^{\alpha_1 + x_i^T \beta}\right) - y_i(\alpha_1 + x_i^T \beta) \tag{4.16}$$

$$\nabla L_s(\overline{\alpha}, \overline{\beta}) = \frac{1}{n_s} \sum_{n=1}^{n_s} (\frac{e^{\overline{\alpha}_s + x_i^T \cdot \overline{\beta}}}{1 + e^{\overline{\alpha}_s + x_i^T \overline{\beta}}} - y_n) x_n \tag{4.17}$$

4.3.1 Initialization

There are 2 options how to initialize the initial coefficients for this method. One is simply set all initial coefficients to 0 as for distributed IRLS method.

The second approach is to estimate the coefficients only using the data from machine 1. This approach provides a reasonable guess of the final coefficients and should allow faster convergence of the algorithm.

4.3.2 Algorithms

The pseudo-code for step by step completion of CSL algorithm is presented in algorithm listing 4.

$\beta_0 = 0$, $\alpha_0 = 0$ or $\alpha_0, \beta_0 = \text{argmin}_{\alpha,\beta} L_1(\alpha, \beta)$ where $\alpha_{0s|s\neq 1} = 0$
$last$ = maximum number of iterations
for $j = 1$ **to** $last$ **do**
  Send α_{j-1}, β_{j-1} to all data nodes
  for $s = 1$ **to** k **do**
    $\nabla L_s(\alpha_{j-1}, \beta_{j-1}) = \frac{1}{n_s} \sum_{n=1}^{n_s} (\frac{e^{\alpha_{j-1,s} + x_i^T \beta_{j-1}}}{1+e^{\alpha_{j-1,s} + x_i^T \beta_{j-1}}} - y_n) x_n$
    Send $\nabla L_s(\alpha_{j-1}, \beta_{j-1})$ to machine 1;
  end
  $\nabla L(\alpha_{j-1}, \beta_{j-1}) = \frac{\sum_{s=1}^{k} \nabla L_s(\alpha_{j-1}, \beta_{j-1})}{k}$
  $\widetilde{L}(\alpha, \beta) = L_1(\alpha, \beta) - (\nabla L_1(\alpha_{j-1}, \beta_{j-1}) - \nabla L(\alpha_{j-1}, \beta_{j-1}))(\alpha \oplus \beta)$;
  $\alpha_j, \beta_j = \text{argmin}_{\alpha,\beta} \widetilde{L}(\alpha, \beta)$;
  if *algorithm converged* **then**
    $last = j$;
  end
end
return $\alpha_{last}, \beta_{last}$

Algorithm 2: CSL algorithm

The algorithm for ODAL is very similar to CSL and presented in algorithm listing 3.

$\overline{\beta} = 0$, $\overline{\alpha} = 0$ or $\overline{\alpha}, \overline{\beta} = \text{argmin}_{\alpha,\beta} L_1(\alpha, \beta), \overline{\alpha}_{s|s\neq 1} = 0$
Send $\overline{\alpha}$, $\overline{\beta}$ to all data nodes
for $s = 1$ **to** k **do**
  $\nabla L_s(\overline{\alpha}, \overline{\beta}) = \frac{1}{n_s} \sum_{n=1}^{n_s} (\frac{e^{\overline{\alpha}_s + x_i^T \overline{\beta}}}{1+e^{\overline{\alpha}_s + x_i^T \overline{\beta}} - y_n) x_n}$
  Send $\nabla L_s(\overline{\alpha}, \overline{\beta})$ to machine 1;
end
$\nabla L(\overline{\alpha}, \overline{\beta}) = \frac{\sum_{s=1}^{k} \nabla L_s(\overline{\alpha}, \overline{\beta})}{k}$
$\widetilde{L}(\alpha, \beta) = L_1(\alpha, \beta) - (\nabla L_1(\overline{\alpha}, \overline{\beta}) - \nabla L(\overline{\alpha}, \overline{\beta}))(\alpha \oplus \beta)$;
$\alpha_{res}, \beta_{res} = \text{argmin}_{\alpha,\beta} \widetilde{L}(\alpha, \beta)$;
return $\alpha_{res}, \beta_{res}$

Algorithm 3: ODAL algorithm

4.4 CSL modification

$\beta_0 = 0$, $\alpha_0 = 0$ or $\alpha_0, \beta_0 = \text{argmin}_{\alpha,\beta} L_1(\alpha, \beta)$ where $\alpha_{0s|s\neq 1} = 0$
$last$ = maximum number of iterations
for $j = 1$ **to** $last$ **do**
  Send α_{j-1}, β_{j-1} to all data nodes
  for $s = 1$ **to** k **do**
    $\nabla L_s(\alpha_{j-1}, \beta_{j-1}) = \frac{1}{n_s} \sum_{n=1}^{n_s} (\frac{e^{\alpha_{j-1,s} + x_i^T \beta_{j-1}}}{1+e^{\alpha_{j-1,s} + x_i^T \beta_{j-1}}} - y_n) x_n$
    Send $\nabla L_s(\alpha_{j-1}, \beta_{j-1})$ to machine 1;
  end
  $\nabla L(\alpha_{j-1}, \beta_{j-1}) = \frac{\sum_{s=1}^{k} \nabla L_s(\alpha_{j-1}, \beta_{j-1})}{k}$
  Send $\nabla L(\alpha_{j-1}, \beta_{j-1})$ to all data nodes
  $\widetilde{L}(\alpha, \beta) = L_1(\alpha, \beta) - (\nabla L_1(\alpha_{j-1}, \beta_{j-1}) - \nabla L(\alpha_{j-1}, \beta_{j-1}))(\alpha \oplus \beta)$;
  $\alpha_j, \beta_j = \text{argmin}_{\alpha,\beta} \widetilde{L}(\alpha, \beta)$;
  if *algorithm converged* **then**
    $last = j$;
  end
end
return $\alpha_{last}, \beta_{last}$

Algorithm 4: CSL modification

Chapter 5

Simulations

5.1 Motivation

To compare the performance of DIALOG methods under different settings we decided to use simulation studies. As described in [27] simulation study is a computer test which uses data generation by pseudo-random sampling taking into account the "true" values of data. As a consequence, simulation studies have an opportunity to detect the performance of statistical methods due to the fact that real data properties are known from the data generation mechanism.

The aims of our simulation study are 1) to find out the effect of simulation parameters on methods performance and 2)to compare estimations from different methods under the same simulation scenario setting. Since the goals are related to the estimation of coefficients, we decided to concentrate only on estimations provided by the methods. Therefore quality of predictions is not assessed since usually it depends on estimated coefficients provided by the methods.

Performance of DIALOG methods would be compared with simple logistic regression which uses the same data combined into one single location. Using the simple logistic regression would allow to detect limitations of distributed logistic regression methods.

5.2 Scenarios

~~Researchers use simulation studies to get empirical results o[illegible]tistical methods performance under particular scenarios. Therefore, each scenario tackles an estimation based on particular goal and this goal specifies inclusion of aspects under which methods need to be tested. As a result, selection of data scenarios is an important aspect and should be highly influenced by the research question.~~

To address the simulation goals ~~and requirements of IKN~~[illegible]e following parameters were selected for the simulation scenarios: 1)number of sites, 2)number of records, 3) different outcome distribution and 4)records distribution among the sites. The following parameters lead to the simulation scenarios described in Ta[illegible]5.1. Additionally, performance of DIALOG methods was also evaluated using the data without a site-specific intercept as shown on scenarios $15 - 17$.

Number	Sites number	Records distribution	Total records	Outcome distribution	Site-specific Intercept	N simulations
1	2	Equal	10000	50/50	Yes	1000
2	2	Equal	10000	95/5	Yes	1000
3	2	Equal	1000	50/50	Yes	1000
4	2	Equal	1000	95/5	Yes	1000
5	4	Equal	10000	50/50	Yes	1000
6	4	Equal	10000	95/5	Yes	1000
7	4	Equal	1000	50/50	Yes	1000

8	4	Equal	1000	95/5	Yes	1000
9	9	Equal	10000	50/50	Yes	1000
10	9	Equal	10000	95/5	Yes	1000
11	9	Equal	1000	50/50	Yes	1000
12	9	Equal	1000	95/5	Yes	1000
13	9	According to table A	10000	50/50	Yes	1000
14	9	According to table A	1000	50/50	Yes	1000
15	9	Equal	10000	50/50	No	1000
16	9	Equal	10000	95/5	No	1000
17	9	According to table A	10000	50/50	No	1000

Table 5.1: Selected scenarios for the simulation

5.2.1 Sites number

Number of sites is an important metric since the DIALOG methods would be used for different research networks. Therefore simulation study should estimate how number of included centers may effect the precision of each of the methods. As mentioned in the Chapter 1, we assume that number of sites would vary from 2 to 9. For that reason simulations were performed for $2, 4$ and 9 sites. These sites have different intercepts when site-specific intercept is present in the data.

We assume that performance of methods when data records are unevenly distributed among sites or when site-specific intercept is not present in the data can be successfully estimated using only 9 sites.

5.2.2 Records distribution

It is an extremely unlikely situations that each of the data sites would have equal number of records in ~~the~~ real life. For example, let consider international country network for studying rare types of cancer. Number of records for each of the countries would usully significantly depends on the country population size, for example, Germany has more than 150 times higher population comparing to Malta. Therefore, approximately the same differences is also likely to meet in the records distribution among sites as well. For that reasons, methods performance was tested for both even and not even data records partioning among sites.

5.2.3 Number of records

Frequency of different type of cancers vary a lot and there are some types which are rare not only in the Netherlands but for the whole hospital network in overall. For that reason, simulations include 2 values for number of records: 1000 and 10000. We assume that total number of records equal to 1000 and 10000 would demonstrate methods performance when 1) sites have a little amount of data records and 2) sites have a sufficient amount of data records accordingly.

5.2.4 Outcome distribution

Outcome variability is one more parameter which would vary in ~~the~~ real life and should be estimated. Different tasks may have different outcome distribution. For example, it is possible that for one type of cancer, 5-years survival rate would be extremely low and for other type it would be more balanced.

Therefore, we need to know if model is able to provide reliable estimates in these situations. Two outcomes distributions were included in the simulation studies. These distributions include ~~the~~ scenarios when positive and negative events are observed approximately for 1)51% and 49%(as in the original data) and 2) 95% and 5% of all data records accordingly. In this simulation study this distribution is controlled by intercepts change.

5.2.5 Site-specific intercept

At the current moment, performance of DIALOG methods was not compared between each other in the literature. To fulfill the gap, there are simulation scenarios which asses methods performance using ~~the~~ data without site-specific intercept. Additionally, this comparison may reveal how site-specific intercept effects the model accuracy and performance.

5.2.6 Number of simulations

To make simulation results reliable and reduce Monte-Carlo Standard Error, a high number of repetitions should be made for each data scenario. We assume that 1000 number of repetitions is sufficient to ensure reliable results

5.3 Data-generating mechanism

For the simulation we generate ~~the~~ data by parametric draw from the pooled logistic regression model ~~which~~ coefficients were estimated based on data set provided by IKNL.

The data set provided by organisation includes three types of variables: numerical, categorical and binary. All numerical variables follow the normal distribution and normalized in order to guarantee ~~the~~ consistent estimates for methods which are independent from parameters of distribution.

Logistic regression requires a creation of dummy variables for both categorical and binary types. When dummy variables are generated one category should be specified as a reference one and this choice for both categorical and binary types of variable can highly influence the final model estimates. For that reason, the category that has the highest number of records was chosen as a reference category.

The full list of original data set ~~variables~~, their frequencies(for binary and categorical data types) and coefficients estimates is provided in the appendix

Put the link to ~~appendix~~

~~The outcome was generated as a random binary variable where probability of outcome equal to 1 was calculated according to formulas 4.2 for scenarios with site-specific intercept and 4.1 for scenarios without site-specific intercept.~~

5.4 Implementation details

To make simulations for different scenarios completely independent from each other, we have generated a random seed for each scenario using *random.choice* function from *numpy* package from Python using preassigned seed equal to 10. These seeds were generated without replacement from the range $0 \dots 150$.

Then the random seed was generated without replacement for each of the repetitions using the same function using the scenario seed. The next step includes generation of a seed without replacement for each of the data variables using the specified repetition seed. The final step is generation of variables inside each of the repetitions using specified seeds. Numerical variables were generated using *random.normal* function, binary and categorical using *random.choice* function.

5.5 Performance measures

As was discussed earlier, the main focus of the simulation is assessment of the estimates quality. For that reason, obtained coefficient's estimated for each method were evaluated using Bias, MAE and MSE and Monte-Carlo Standard Error measures. These measures were calculated according to formulas 5.1, 5.2, 5.3 and 5.4.

$$Bias_c = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{i,c} - \beta_c) \tag{5.1}$$

$$MAE_c = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{\beta}_{i,c} - \beta_c\right| \tag{5.2}$$

$$MSE_c = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{i,c} - \beta_c)^2 \tag{5.3}$$

$$Monte - Carlo\ SE_c = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\beta}_{i,c} - \bar{\beta}_c)^2} \tag{5.4}$$

where c means a data covariate, n means number of repetitions for simulation scenario, i means a repetition, $\hat{\beta}_{i,c}$ means a model estimate for covariate c in repetition i, β_c means a coefficient ~~which is used~~ for covariate c during data simulation(real coefficient value) and $\bar{\beta}_c$ means an average model estimate for covariate c for all repetitions.

Additionally, for allowing ~~the~~ methods comparison using the single value, simulation MSE was proposed which is calculated according to formula 5.5.

$$Simulation\ MSE = \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{d}(\hat{\beta}_{i,c} - \beta_c)^2 \tag{5.5}$$

where d is an overall number of data covariates used for the simulation scenario i.

Bibliography

[1] Data explorer — ECIS. 1

[2] Data explorer — ECIS. 2

[3] Paul D. Allison. Convergence failures in logistic regression. 2008. 6, 7, 8

[4] Yoshinori AONO, Takuya HAYASHI, Le Trieu PHONG, and Lihua WANG. Privacy-Preserving Logistic Regression with Distributed Data Sources via Homomorphic Encryption. *IEICE Transactions on Information and Systems*, E99.D(8):2079–2089, aug 2016. 10

[5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2010. 9

[6] Xi Chen, Weidong Liu, and Yichen Zhang. First-order Newton-type Estimator for Distributed Estimation and Inference. nov 2018. 9

[7] Wenrui Dai, Shuang Wang, Hongkai Xiong, and Xiaoqian Jiang. Privacy Preserving Federated Big Data Analysis. pages 49–82. Springer, Cham, 2018. 1, 3

[8] Timo M. Deist, A. Jochems, Johan van Soest, Georgi Nalbantov, Cary Oberije, Seán Walsh, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Andre Dekker, and Philippe Lambin. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and Translational Radiation Oncology*, 4:24–31, jun 2017. 5

[9] Wei Du, Ang Li, and Qinghua Li. Privacy-Preserving Multiparty Learning For Logistic Regression. oct 2018. 10

[10] Rui Duan, Mary Regina Boland, Jason Moore, and Yong Chen. Odal: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 24:30–41, 01 2019. 13

[11] Khaled El Emam, Saeed Samet, Luk Arbuckle, Robyn Tamblyn, Craig Earle, and Murat Kantarcioglu. A secure distributed logistic regression protocol for the detection of rare adverse drug events. *Journal of the American Medical Informatics Association*, 20(3):453–461, may 2013. 12

[12] Stephen E. Fienberg, Yuval Nardi, and Aleksandra B. Slavković. Valid Statistical Analysis for Logistic Regression with Multiple Sources. pages 82–94. Springer, Berlin, Heidelberg, 2009. 10, 12

[13] Amadou Gaye, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, Andrew Turner, Elinor M Jones, Joel Minion, Andrew W Boyd, Christopher J Newby, Marja-Liisa Nuotio, Rebecca Wilson, Oliver Butters, Barnaby Murtagh, Ipek Demir, Dany Doiron, Lisette Giepmans, Susan E Wallace, Isabelle Budin-Ljøsne, Carsten Oliver Schmidt, Paolo Boffetta, Mathieu

Boniol, Maria Bota, Kim W Carter, Nick DeKlerk, Chris Dibben, Richard W Francis, Tero Hiekkalinna, Kristian Hveem, Kirsti Kvaløy, Sean Millar, Ivan J Perry, Annette Peters, Catherine M Phillips, Frank Popham, Gillian Raab, Eva Reischl, Nuala Sheehan, Melanie Waldenberger, Markus Perola, Edwin van den Heuvel, John Macleod, Bartha M Knoppers, Ronald P Stolk, Isabel Fortier, Jennifer R Harris, Bruce HR Woffenbuttel, Madeleine J Murtagh, Vincent Ferretti, and Paul R Burton. DataSHIELD: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology*, 43(6):1929–1944, dec 2014. 5

[14] Rob Hall, Yuval Nardi, and Stephen Fienberg. Achieving Both Valid and Secure Logistic Regression Analysis on Aggregated Data from Different Private Sources. nov 2011. 10

[15] David W. Hosmer, Stanley. Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression.* 5, 6

[16] Yichen Jiang, Jenny Hamer, Chenghong Wang, Xiaoqian Jiang, Miran Kim, Yongsoo Song, Yuhou Xia, Noman Mohammed, Md Nazmus Sadat, and Shuang Wang. SecureLR: Secure Logistic Regression Model via a Hybrid Cryptographic Protocol. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(1):113–123, jan 2019. 10

[17] Elinor M. Jones, Nuala A. Sheehan, Amadou Gaye, Philippe Laflamme, and Paul Burton. Combined analysis of correlated data when data cannot be pooled. *Stat*, 2(1):72–85, dec 2013. 12

[18] E.M. Jones, N.A. Sheehan, N. Masca, S.E. Wallace, M.J. Murtagh, and P.R. Burton. DataSHIELD shared individual-level analysis without sharing the data: a biostatistical perspective. *Norsk Epidemiologi*, 21(2), apr 2012. 10, 11

[19] Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-Efficient Distributed Statistical Inference. *Journal of the American Statistical Association*, pages 1–14, feb 2018. 12

[20] Dongyeop Kang, Woosang Lim, Kijung Shin, Lee Sael, and U. Kang. Data/Feature Distributed Stochastic Coordinate Descent for Logistic Regression. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*, pages 1269–1278, New York, New York, USA, 2014. ACM Press. 9

[21] Andrey Kim, Yongsoo Song, Miran Kim, Keewoo Lee, and Jung Hee Cheon. Logistic regression model training based on the approximate homomorphic encryption. *BMC Medical Genomics*, 11(S4):83, oct 2018. 10

[22] Miran Kim, Yongsoo Song, Shuang Wang, Yuhou Xia, and Xiaoqian Jiang. Secure Logistic Regression Based on Homomorphic Encryption: Design and Evaluation. *JMIR medical informatics*, 6(2):e19, apr 2018. 10

[23] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. oct 2016. 10

[24] Wenfa Li, Hongzhe Liu, Peng Yang, and Wei Xie. Supporting Regularized Logistic Regression Privately and Efficiently. *PLOS ONE*, 11(6):e0156479, jun 2016. 12

[25] Qing Ling, Wei Shi, Gang Wu, and Alejandro Ribeiro. DLM: Decentralized Linearized Alternating Direction Method of Multipliers. *IEEE Transactions on Signal Processing*, 63(15):4051–4064, aug 2015. 10

[26] Aryan Mokhtari, Wei Shi, Qing Ling, and Alejandro Ribeiro. A Decentralized Second-Order Method with Exact Linear Convergence Rate for Consensus Optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(4):507–522, dec 2016. 10

[27] Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, may 2019. 21

[28] L. Ohno-Machado, Z. Agha, D. S. Bell, L. Dahm, M. E. Day, J. N. Doctor, D. Gabriel, M. K. Kahlon, K. K. Kim, M. Hogarth, M. E. Matheny, D. Meeker, J. R. Nebeker, F. Resnic, D. Khodyakov, L. Armstead, T. Nagler, S. Morley, N. Anderson, D. Cooper, D. Phillips, D. Heber, Z. Li, M. K. Ong, A. Patel, M. Zachariah, J. C. Burns, L. B. Daniels, S. Doan, C. Farcas, R. Germann-Kurtz, X. Jiang, H.-e. Kim, P. Paul, H. Taras, A. Tremoulet, S. Wang, W. Zhu, D. Berman, A. Rizk-Jackson, M. D'Arcy, C. Kesselman, T. Knight, L. Pearlman, P. Heidenreich, D. Rifkin, C. Stepnowsky, T. Zamora, S. L. DuVall, L. J. Frey, J. Scehnet, B. C. Sauer, J. C. Facelli, R. K. Gouripeddi, J. Denton, F. FitzHenry, J. Fly, V. Messina, F. Minter, L. Nookala, H. Sullivan, T. Speroff, and D. Westerman. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. *Journal of the American Medical Informatics Association*, 21(4):621–626, jul 2014. 5

[29] Zebang Shen, Aryan Mokhtari, Tengfei Zhou, Peilin Zhao, and Hui Qian. Towards More Efficient Stochastic Decentralized Learning: Faster Convergence and Sparse Communication. may 2018. 10

[30] Haoyi Shi, Chao Jiang, Wenrui Dai, Xiaoqian Jiang, Yuzhe Tang, Lucila Ohno-Machado, and Shuang Wang. Secure Multi-pArty Computation Grid LOgistic REgression (SMAC-GLORE). *BMC Medical Informatics and Decision Making*, 16(S3):89, jul 2016. 12

[31] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization. *SIAM Journal on Optimization*, 25(2):944–966, jan 2015. 10

[32] Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient Distributed Learning with Sparsity. may 2016. 12

[33] Shuang Wang, Yuchen Zhang, Wenrui Dai, Kristin Lauter, Miran Kim, Yuzhe Tang, Hongkai Xiong, and Xiaoqian Jiang. HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS. *Bioinformatics*, 32(2):btv563, oct 2015. 10

[34] Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, Isabel Fortier, Bartha M Knoppers, and Paul R Burton. DataSHIELD: resolving a conflict in contemporary bioscience–performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5):1372–82, oct 2010. 4, 5, 10

[35] Yuan Wu, Xiaoqian Jiang, Jihoon Kim, and Lucila Ohno-Machado. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *Journal of the American Medical Informatics Association : JAMIA*, 19(5):758–64, 2012. 11, 12

[36] Wei Xie, Yang Wang, Steven M. Boker, and Donald E. Brown. PrivLogit: Efficient Privacy-preserving Logistic Regression by Tailoring Numerical Optimizers. nov 2016. 12

[37] Xu Dong Zhu, Hui Li, and Feng Hua Li. Privacy-preserving logistic regression outsourcing in cloud computing. *International Journal of Grid and Utility Computing*, 4(2/3):144, 2013. 10