Computing Generalized Linear Model using Iteratively Weighted Least Squares and Coordinate Descent

Khoa Huynh
Advisor: Dr. Xia Wang
Department of Mathematical Sciences
Division of Statistics and Data Science
University of Cincinnati

August 31, 2020

Abstract

Generalized Linear Model (glm) has been widely used in regression models. Iteratively weighted least squares algorithm (IWLS) and coordinate descent (CD) algorithm are technique that can used to find maximum likelihood function for generalized linear model. We are interested in comparing the two algorithm to the framework of regression models with binary responses. Link function is an important component in the generalized linear model. We carried out a simulation study to compare glm using IWLS algorithm and CD algorithm performance under different link functions. The link functions we investigated include the commonly used logistic (logit), probit and complementary log-log links (cloglog). We also wrote an algorithm glm using iteratively weighted least squares and coordinate descent algorithm. Our results show that under IWLS algorithm perform better than glm under CD algorithm when we increase the number of variable. The current study helps our future research to build an integrated process of variable selection via lasso along with a flexible link function.

Keywords: binary response, coordinate descent algorithm, iteratively weighted least squares, link functions, generalized linear model.

1 Introduction

Generalized linear models include a link function that provides the relationship between linear predictor and the mean of the response random variable. Iteratively weighted least squares is one of the original technique for solving the generalized linear models problems[4]. However, coordinate descent algorithm solve optimization problems and have been used in applications many years. They are iterative methods in which each iterate is obtained by fixing most components of the variable vector at their values from the current iteration, and approximately minimizing the objective with respect to the remaining components[5]. There has been an enormous amount of research activity related to computational algorithm on generalized linear models using coordinate descent such as:

- Coordinate descent algorithms (Stephen J. Wright 2015) describes the fundamentals of the coordinate descent approach on glm, together with variants and extensions and their convergence properties, mostly with reference to convex objectives.
- Coordinate Descent (Geoff Gordon and Ryan Tibshirani 2012) represent the application of coordinate descent algorithms in least absolute shrinkage and selection operators (lasso).
- Coordinate Descent Optimization for l^1 Minimization with Application to Compressed Sensing; a Greedy Algorithm (Yingying Li and Stanley Osher) propose a fast algorithm for solving the Basis Pursuit problem, which has application to compressed sensing.

In this paper, we discuss about the compare accuracy between IWLR algorithm and CD algorithm with link functions such as probit, logit and cloglog, and then compare the time performance of coordinate descent with iteratively weighted least squares. The algorithm iteratively weighted least squares for maximize likelihood estimates function for generalized linear models will be provides (section 2.3) as well as coordinate descent algorithm under probit, logit, and cloglog link functions (section 2.4).

2 Methods

2.1 Generalized Linear Models

Let consider we have binary response variables $Y_1, Y_2, ...Y_n$, and an $n \times p$ matrix of predictor X, where $X = (X_1', X_2', \ldots, X_n')'$ and X_i is the p-dimensional

row vector of predictors, and we want to model the probability $p = Pr(Y = 1)$. The link function can take continous scales from negative infinity to positive infinity and transforms it to the probabilities with value between 0 and 1 for binary response such as logit, probit and cloglog link functions. Since our approach is studying different link functions, it is more convenience to express the logit, probit and cloglog model in terms of function form below:

Table 1: Binary link functions

Link name	Link Function, $\boldsymbol{X}\beta = g(p)$	Inverse Link
Logit	$\boldsymbol{X}\beta = \log(\frac{p}{1-p})$	$p = \frac{\exp(\boldsymbol{X}\beta)}{1+\exp(\boldsymbol{X}\beta)}$
Probit	$\boldsymbol{X}\beta = \Phi^{-1}(p)^*$	$p = \Phi(\boldsymbol{X}\beta)$
Cloglog	$\boldsymbol{X}\beta = \log(-\log(1-p))$	$p = 1 - \exp(-\exp(\boldsymbol{X}\beta))$

Note*: $\Phi(.)$ denote the standard cumulative distribution function of the class, such as the $N(0, 1)$.

We now define the maximum likelihood for function. Let $p(x_i) = Pr(Y = 1|x_i)$ be a probability for observation i at a particular value for the parameters (β_0, β):

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}. \tag{1}$$

The log-likelihood function turns the products into the sums:

$$\ln L(\beta_0, \beta) = \sum_{i=1}^{n} y_i \ln p(x_i) + (1 - y_i) \ln(1 - p(x_i)). \tag{2}$$

2.2 Log maximum likelihood for link functions

1. Logit link function model:

$$\ln L(\beta_0, \beta) = \sum_{i=1}^{N} (y_i \ln(\frac{\exp(\beta_0 + x_i^T\beta)}{1 + \exp(\beta_0 + x_i^T\beta)}) + (1-y_i) \ln(1 - \frac{\exp(\beta_0 + x_i^T\beta)}{1 + \exp(\beta_0 + x_i^T\beta)})) \tag{3}$$

2. Probit link function model:

$$\ln L(\beta_0, \beta) = \sum_{i=1}^{N} (y_i \ln(\Phi(\beta_0 + x_i^T\beta)) + (1 - y_i) \ln(1 - \Phi(\beta_0 + x_i^T\beta))) \tag{4}$$

3. Cloglog link function model:

$$\ln L(\beta_0, \beta) = \sum_{i=1}^{N} (y_i \ln(1 - \exp(-\exp(\beta_0 + x_i^T \beta))) + (1 - y_i)(-\exp(\beta_0 + x_i^T \beta))). \tag{5}$$

2.3 Iteratively Weighted Least Squares

Iteratively Weighted Least Squares (IWLS) is used to find the maximum likelihood estimates of a generalized linear model [4]. The following table shows the algorithm for IWLS, where r^{th} is the number of iteration :

Step 1: Let β^r the current estimate of $\hat{\beta}$, determine
$\hat{\eta_i^r} := x_i^T \beta^r, i = 1, ..., n$, (current linear predictor)
$\hat{\mu_i^r} := g^{-1}(\eta_i^r)$, (current fitted means)
$\nu_i^r := a(\phi) b"(\theta_i)|_{\theta_i = \hat{\theta_i^r}}$
$\hat{\theta_i^r} := h^{-1}(\mu_i^r)$
$Z_i^r := \hat{\eta_i^r} + (y_i - \hat{\mu_i^r})(\frac{d\eta_i}{d\mu_i}|_{\eta_i = \hat{\eta_i^r}})$, (adjusted dependent variable)
$W_i^r := [\nu_i^r (\frac{d\eta_i}{d\mu_i}|_{\eta_i = \hat{\eta_i^r}})^2]^{-1}$
Step 2: Regress Z_i^r on $x_{i1}, x_{i2}, ...x_{ip}$ with weights $(W_i^r)^{-1}$ to obtain new estimate β^{r+1} and continue with step 1 until $|\beta^r - \beta^{r+1}|$ sufficiently small.

2.4 Coordinate Descent

Coordinate descent algorithms is an optimization algorithm that minimizes along the coordinate direction to find the minimum function. Coordinate descent has been applied to several algorithms for solving the optimization problem. The idea behind these algorithms is simple. Suppose f is multiple dimensional functions, in our case it would be the log-likelihod function. We can minimize f by minimizing each of the individual dimensions of f, while holding the argument of f in the other dimensions fixed [3]. In other word, it will take complex multiple dimensional problems and reduces it to a collection of one-dimensional problem [3].
While we keeping $\beta_0 = 0$ and moving β_1, we can see in Fig 2 that the logit likelihood have the minimum to achieve.

Here is the write up algorithm for coordinate:

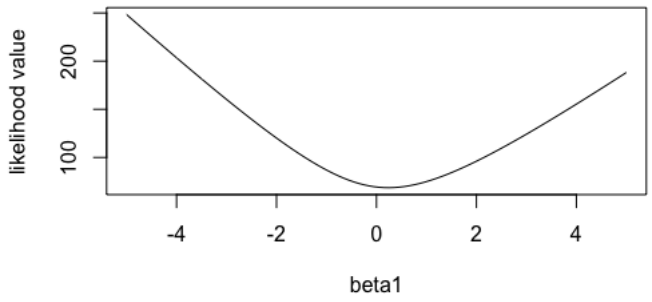

Figure 1: One-Dimensional function to optimize

Step 1:
Set initial point $\beta = (0, ...0)$ Calculate $\ln L(0, ...0)^r$ the current estimate of log-likelihood Step 3:
For $j = 1, ..., p$, minimize $\ln L(\beta_0, ..., \beta_{j-1}, \beta, \beta_{j+1}, ...\beta_p)^{r+1}$ where $\beta_0, ..., \beta_{j-1}, \beta_{j+1}, ...\beta_p$ are all fixed at their current values.
Step 4:
Continue with step 3 until $|\ln(.)^r - \ln(.)^{r+1}|$ sufficiently small

3 Simulation Study

3.1 Example 1

In this example, we investigate on the comparison between IWLS and CD algorithm. We used $\beta = (1, 3)$ to generate data from multivariate normal distribution $\boldsymbol{X_i}$, where $i = 1, 2$. The data consist of response y variables from the model:

$$g(E(Y_i)) = \sum_{j=1}^{p} \beta_j \boldsymbol{X_{ij}}$$

. We run over 100 simulations. The correlation between x_i and x_j, $j = 1, 2..., 1000$, was $\rho^{|i-j|}$ with $\rho = 0.90$. We apply logit, probit and cloglog link function as the true model in each simulations. We split data into 2 parts: train data and test data base on the ratio 1:1. We used train data to find $\hat{\beta}$, which is the estimate of β and then using $\hat{\beta}$ with test data to compute the response. After that, we calculate the accuracy of the test data for each methods (IWLS and CD).
Here is the result for simulation comparision between IWLS and CD with

logit link functions:

Table 2: Accuracy for Example 1

	Logit		Probit		Cloglog	
	IWLS	CD	IWLS	CD	IWLS	CD
Mean	78%	77.95%	86.3%	86.57%	76.01%	85.41%
Median	78.2%	78%	86.2%	86.7%	75.%	85.50%

As we can see from the table 2, the mean and median accuracy from IWLS methods slightly higher than CD methods for logit link. However, CD methods is better than IWLS method for both probit and cloglog link function.
The boxplot below show time comparision between IWLS and CD with logit link functions. Base on the boxplot, the time to converge for CD method is faster compare to IWLS methods. This is the same case for both probit and cloglog link functions (graph in appendix section).

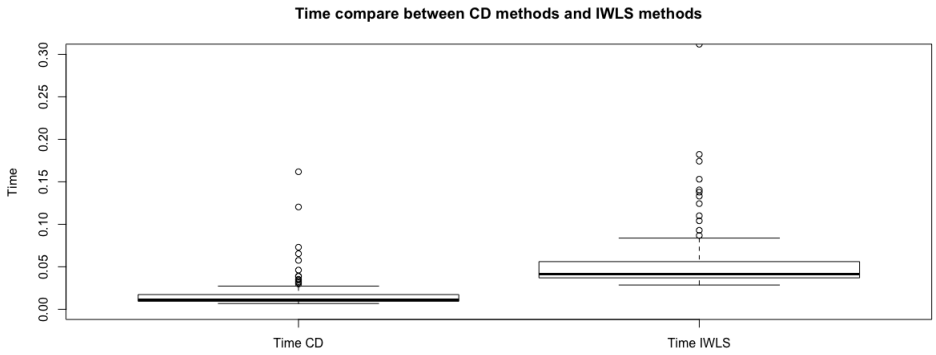

Figure 2: Boxplot time comparision

3.2 Example 2

In this example, we used the same set up as we did in example 1, however, with $\beta = (1, -3, 3)$ generate data from multivariate normal distribution $\boldsymbol{X_i}$, where $i = 1, 2, 3$. The correlation between x_i and x_j, $j = 1, 2..., 1000$, was $\rho^{|i-j|}$ with $\rho = 0.99$ for example 2a, $\rho = 0.01$ for example 2b,$\rho = 0.5$ for

example 2c . The data consist of response y variables from the model:

$$g(E(Y_i)) = \sum_{j=1}^{p} \beta_j \boldsymbol{X_{ij}}$$

Here is the result for simulation comparision between IWLS and CD with logit, probit and cloglog link functions:

Table 3: Accuracy for Example 2a

	Logit		Probit		Cloglog	
	IWLS	CD	IWLS	CD	IWLS	CD
Mean	61.51%	61.29%	72.76%	72.37%	85.33%	85.06%
Median	61.30%	61.40%	73.00%	72.40%	85.40%	85.40%

Table 4: Accuracy for Example 2b

	Logit		Probit		Cloglog	
	IWLS	CD	IWLS	CD	IWLS	CD
Mean	82.96%	83.21%	89.92%	89.82%	77.80%	84.36%
Median	83%	83.20%	90.10%	89.80%	78.00%	86.00%

Table 5: Accuracy for Example 2c

	Logit		Probit		Cloglog	
	IWLS	CD	IWLS	CD	IWLS	CD
Mean	77.71%	77.90%	86.43%	86.35%	78.33%	83.06%
Median	78.00%	78.00%	86.50%	86.40%	79.40%	84.42%

As we can see, the table show that both mean and median accuracy for IWLS and CD method are the same. However, the CD method perform better in example 2b and example 2c than IWLS method.

The boxplot from figure 3 and figure 5 show time comparision between IWLS and CD with logit link functions for example 2a and example 2c. Time convergence for CD methods about 10 higher than time convergence for IWLS methods. However, the time convergence for CD method in figure 4 (example 2b) lower than time convergence for IWLS method. Base on boxplots below, we can see that the correlation have effect on the CD algorithm.

This is the same case for both probit and cloglog link functions (graph in appendix section).

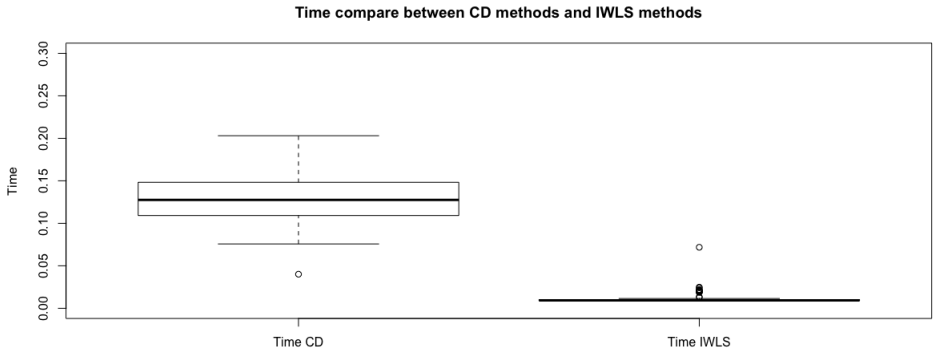

Figure 3: Example 2a: Boxplot time comparision

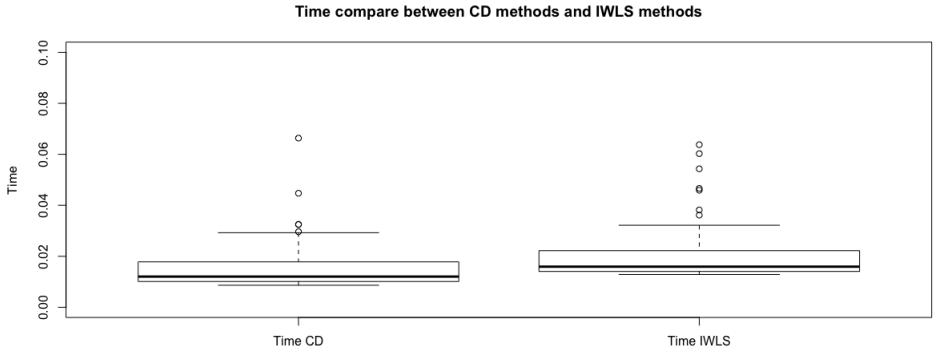

Figure 4: Example 2b: Boxplot time comparision

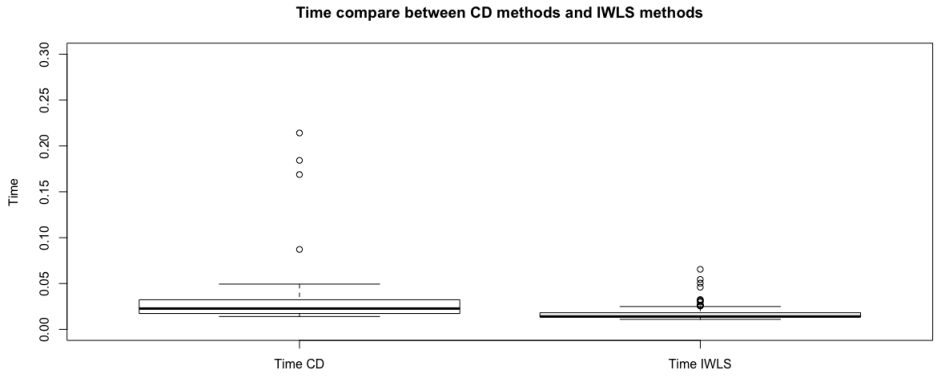

Figure 5: Example 2c: Boxplot time comparision

3.3 Convergence

The correlation for data does not impact on the CD algorithm as we can see in figure 8, which is $\rho = 0.01$. Moreover, the time convergence for IWLS are higher than CD method like we discuss in figure 2. Both figure 6, figure 7 and figure 9 show that the higher correlation, the more time CD algorithm need to convergence. CD algorithm convergence much faster and stable with variable does not have correlation.

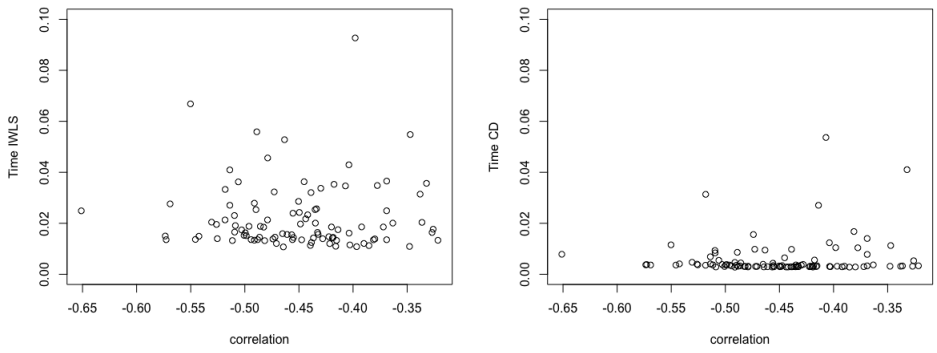

Figure 6: Example 1 - Correlation verse time between IWLS and CD for logit link

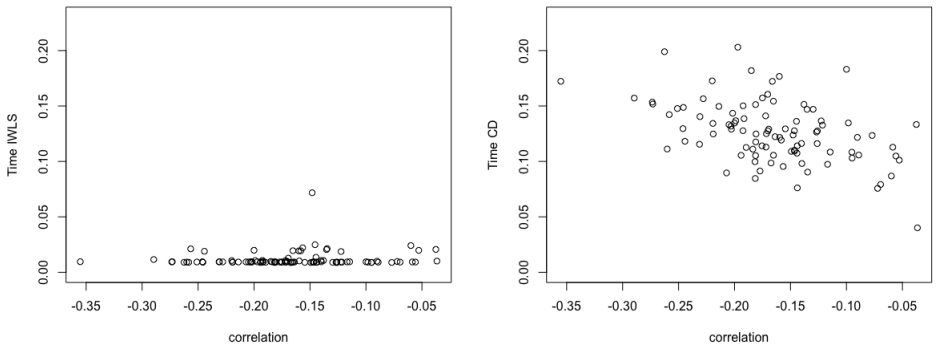

Figure 7: Example 2a - Correlation verse time between IWLS and CD for logit link

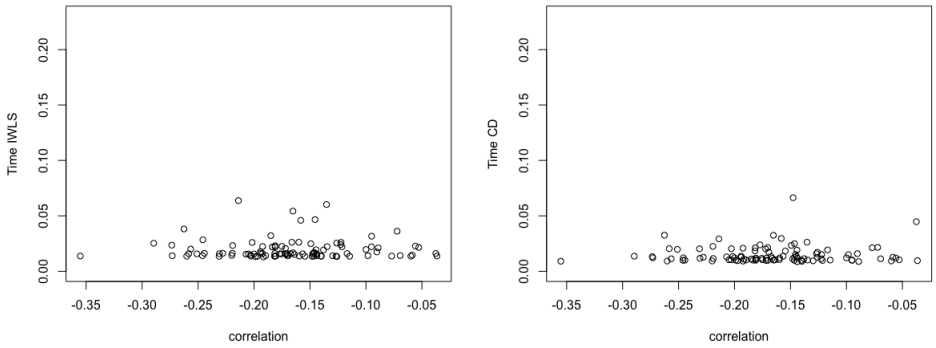

Figure 8: Example 2b - Correlation verse time between IWLS and CD for logit link

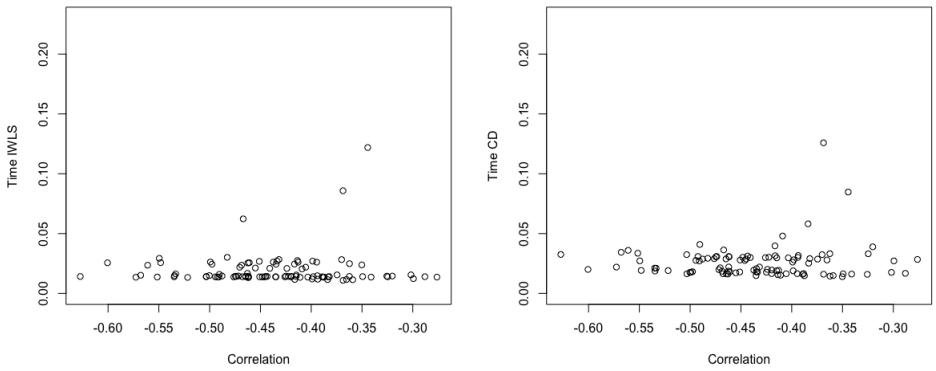

Figure 9: Example 2c - Correlation verse time between IWLS and CD for logit link

4 Discussion and Future Works

In the simulation study, although the model under misspecfiy link function, the CD algorithm have better performance accuarcy more than IWLS algorithm for glm. Moreover, lasso model have very high percentage to selecting the correct variable under different link function. The misspecified link functions does not effect on the shrinkage methods for choosing the correct variable. For the future work, we would like to continous investiagte on the link function with difference misspecify approach to see whether the true link function can have higher accuracy than other link functions.

A Graph for probit link simulation

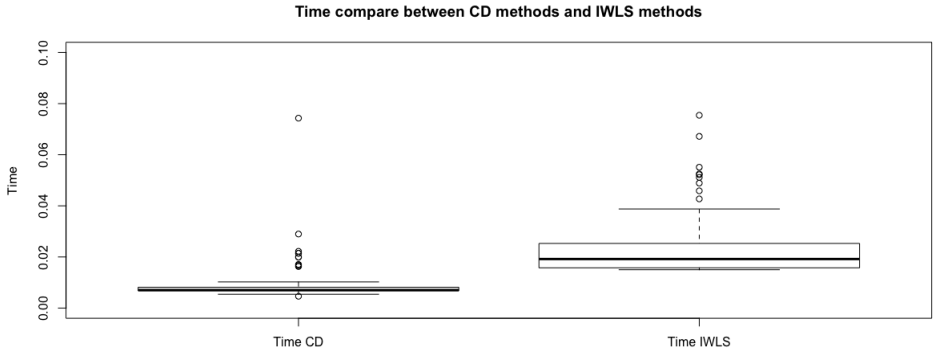

Figure 10: Example 1: Boxplot time comparision probit link

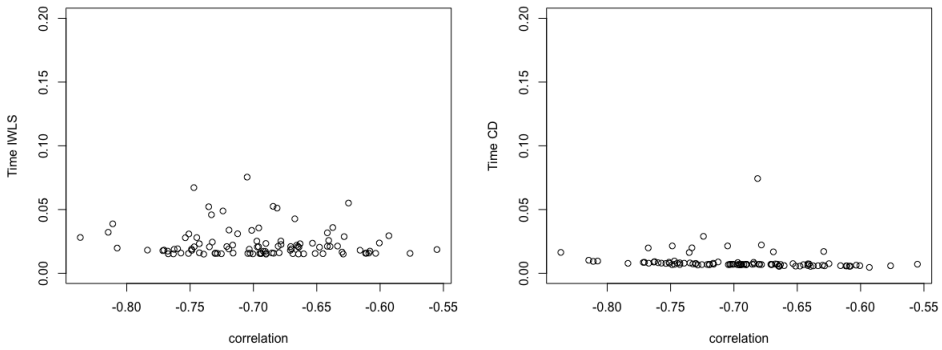

Figure 11: Example 1: Correlation verse time between IWLS and CD for probit link

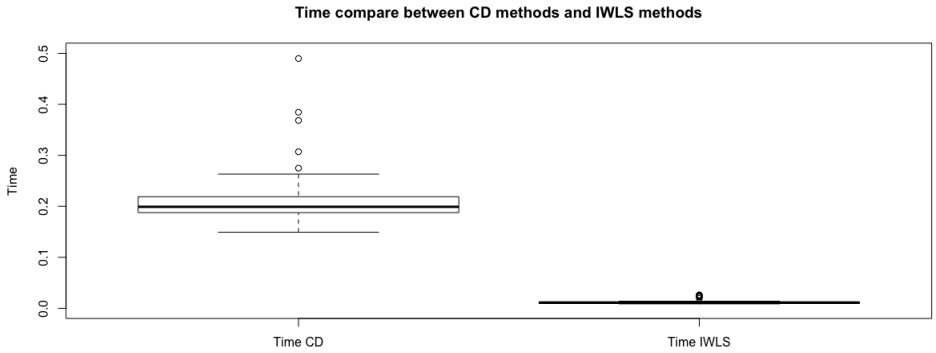

Figure 12: Example 2a: Boxplot time comparision probit link

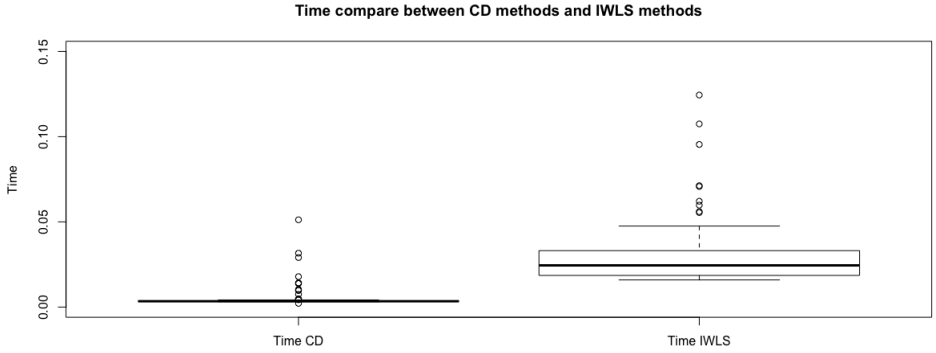

Figure 13: Example 2b: Boxplot time comparision probit link

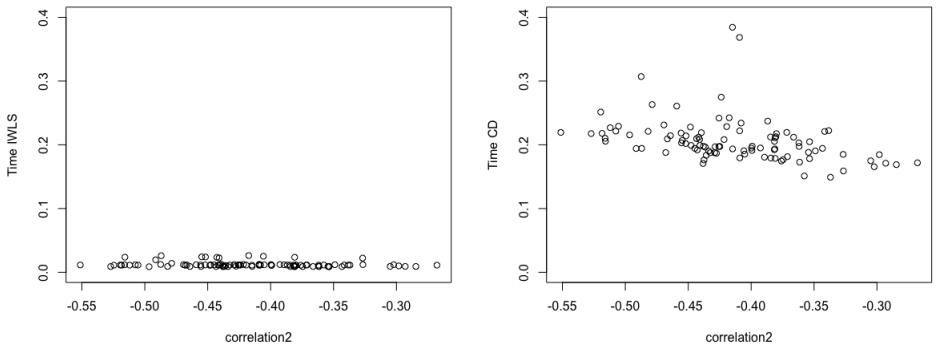

Figure 14: Example 2a: Correlation verse time between IWLS and CD for probit link

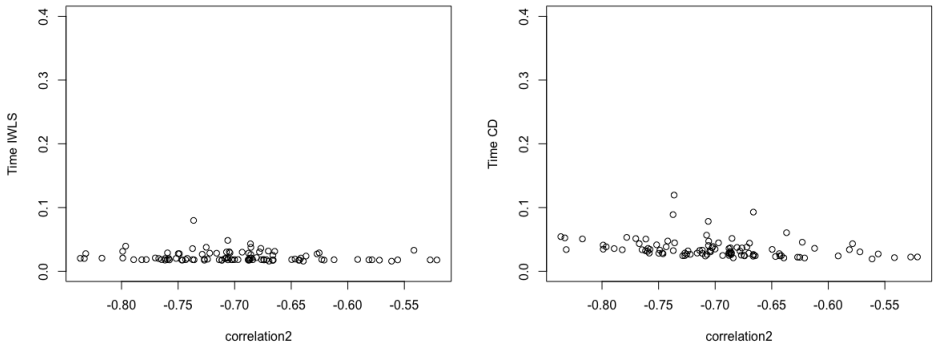

Figure 15: Example 2b: Correlation verse time between IWLS and CD for probit link

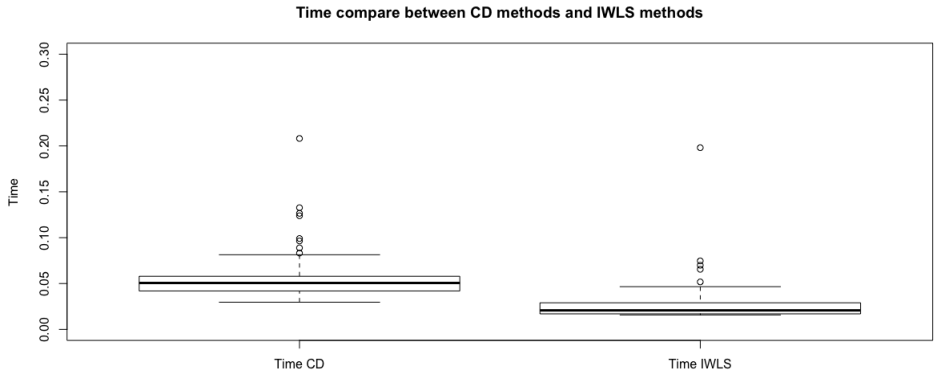

Figure 16: Example 2c: Boxplot time comparision probit link

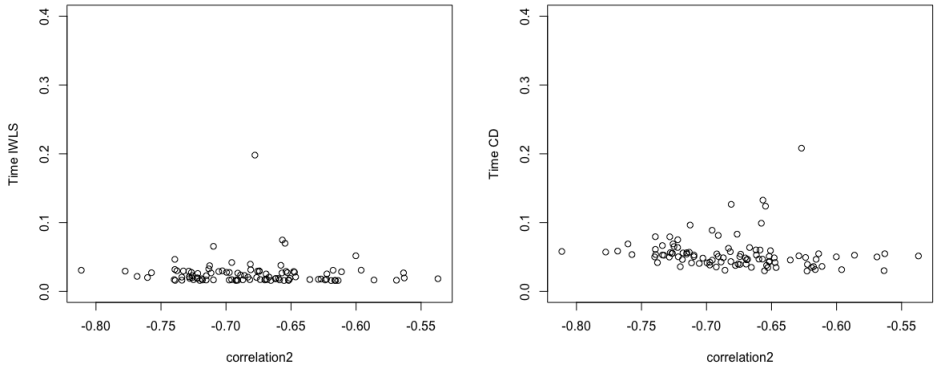

Figure 17: Example 2c: Correlation verse time between IWLS and CD for probit link

B Graph for cloglog link simulation

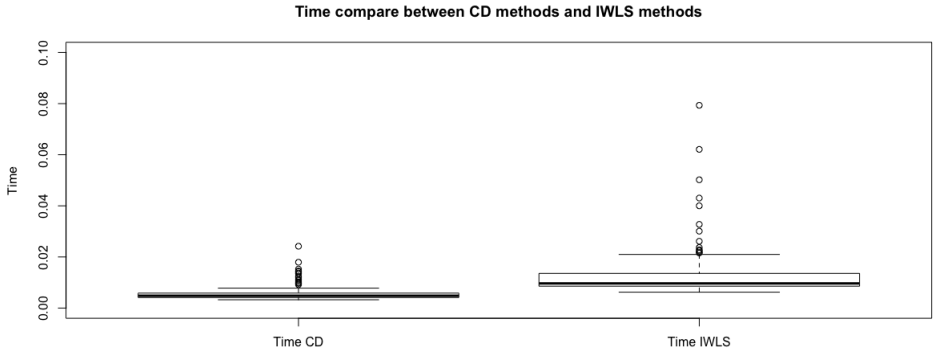

Figure 18: Example 1: Boxplot time comparision cloglog link

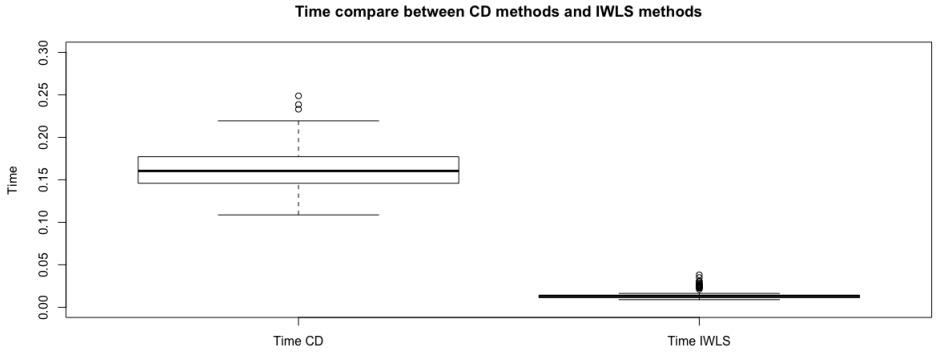

Figure 19: Example 2a: Boxplot time comparision cloglog link

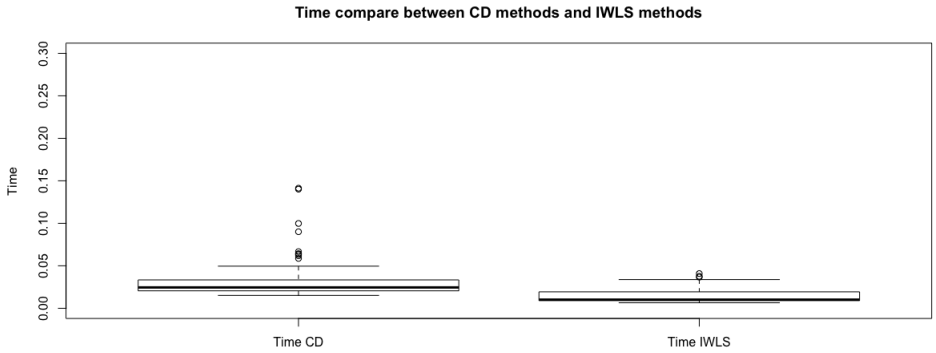

Figure 20: Example 2b: Boxplot time comparision cloglog link

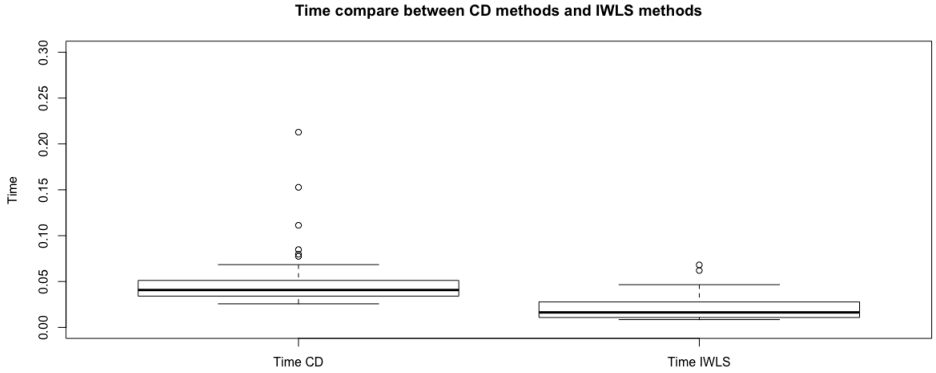

Figure 21: Example 2c: Boxplot time comparision cloglog link

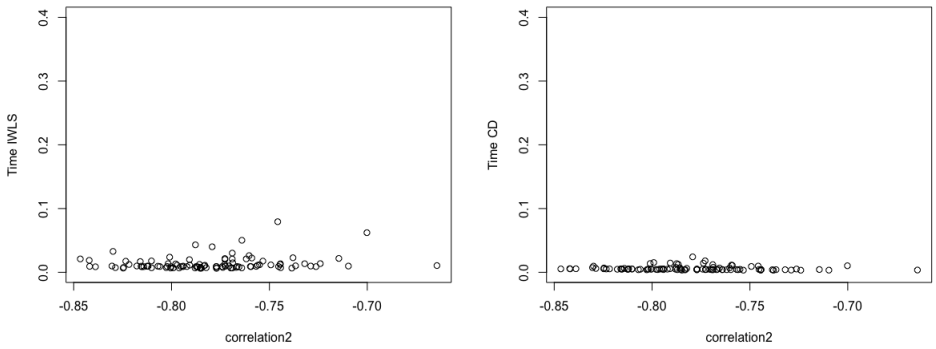

Figure 22: Example 1 - Correlation verse time between IWLS and CD for cloglog link

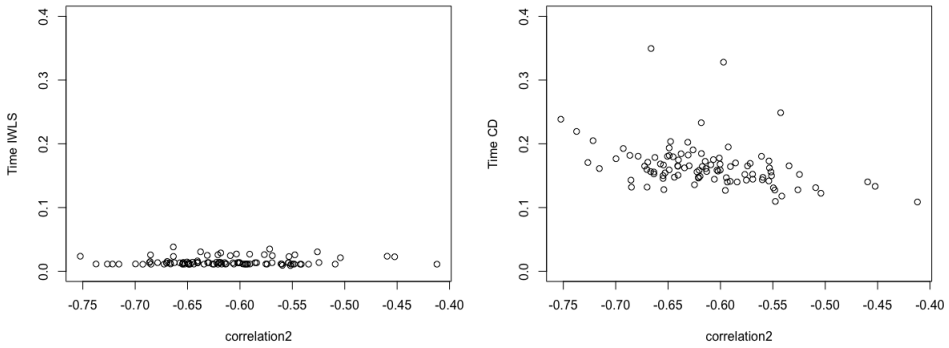

Figure 23: Example 2a - Correlation verse time between IWLS and CD for cloglog link

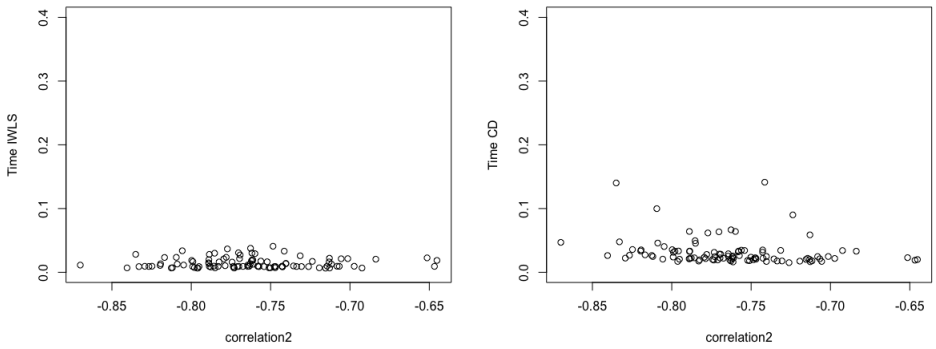

Figure 24: Example 2b - Correlation verse time between IWLS and CD for cloglog link

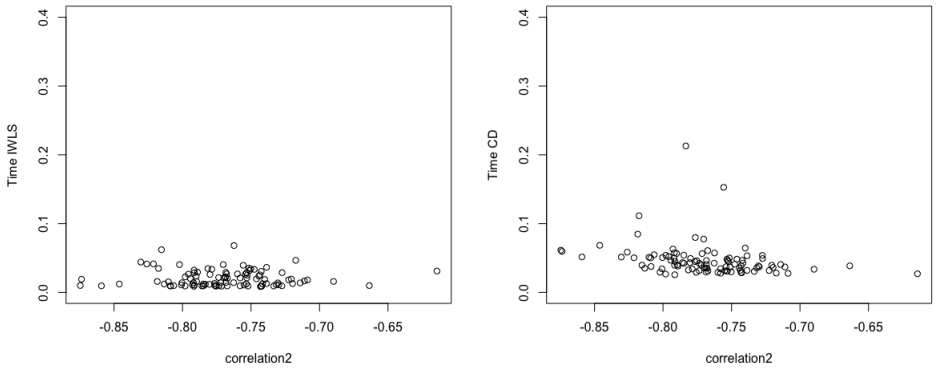

Figure 25: Example 2c - Correlation verse time between IWLS and CD for cloglog link

References

[1] Jerome Friedman, Trevor Hastie, and Rob Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent", Journal of Statistical Software, vol. 33, Issue 1, January 2010.

[2] Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso",Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, No. 1, pp. 267-288, 1996.

[3] Robert D. Peng, "Advanced Statistical Computing" 17 July 2018, https://bookdown.org/rdpeng/advstatcomp/ikelihood.png

[4] Nelder, J. A. and Wedderburn, R. W. M. "Generalized linear models." J.R.Statist. Soc.A, 135,370-384. 1972

[5] Wright, S.J. "Coordinate descent algorithms". Math. Program. 151, 3–34 (2015). https://doi.org/10.1007/s10107-015-0892-3

[6] Geoff Gordon, Ryan Tibshirani. "Coordinate Descent." 2012, https://www.cs.cmu.edu/ ggordon/10725-F12/slides/25-coord-desc.pdf

[7] Yingying Li, Stanley Osher. "Coordinate Descent Optimization for l^1 Minimization with Application to Compressed Sensing; a Greedy Algorithm" 2009. ftp://ftp.math.ucla.edu/pub/camreport/cam09-17.pdf