

Carnegie Mellon University

CARNEGIE INSTITUTE OF TECHNOLOGY

REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Master of Science

TITLE Techniques for Group-Wise Feature Selection and Estimation

PRESENTED BY Aditya Chindhade

ACCEPTED BY THE DEPARTMENT OF

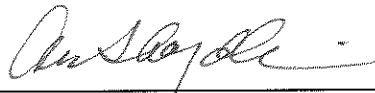
Chemical Engineering



NIKOLAOS V. SAHINIDIS, PROFESSOR AND ADVISOR

DATE

12/3/18



ANNE S. ROBINSON, DEPARTMENT HEAD

DATE

12/6/2018

Techniques for group-wise feature selection and estimation

Aditya Chindhade
Carnegie Mellon University
Master's Project Report
2018

Advisor: Prof. Nikolaos V. Sahindis

Contents

1	Abstract	1
2	Introduction	2
3	Regularization and feature selection	3
3.1	The Ridge	3
3.2	The Lasso	4
3.3	The Elastic Net	5
3.4	The Group-Lasso	6
3.5	Best subset selection	7
4	ALAMO (Automated Learning of Algebraic Models for Optimization)	8
4.1	The ALAMO model-building process	8
4.1.1	Surrogate Model Generation	8
4.1.2	Adaptive Sampling	8
4.2	Constrained regression in ALAMO	8
5	Experiments	10
5.1	Dataset description	10
5.2	Feature transformations	10
5.2.1	Feature grouping	10
5.3	Correlation plot	11
5.4	Modeling	12
5.4.1	Generalized linear models	12
5.4.2	ALAMO	12
6	Results and Discussion	15
6.1	Linear Regression	15
6.2	Ridge	16
6.3	Lasso	18
6.4	Elastic Net	20
6.5	Group Lasso	22
6.6	ALAMO	24
7	Conclusion	25

List of Figures

1	The ridge penalty as a constrained minimization problem	3
2	The lasso penalty as a constrained minimization problem	4
3	The elastic net penalty as a constrained minimization problem	5
4	The group lasso penalty as a constrained minimization problem	6
5	The best-subset selection technique	7
6	The ALAMO approach	9
7	Identification of feature groups	11
8	Correlation plot: Pearson correlation of each feature with target variable . .	11
9	Ridge coefficients vs base-10 log of tuning parameter	16
10	Ridge mean-squared error vs base-10 log of tuning parameter	17
11	Lasso coefficients vs base-10 log of tuning parameter	18
12	Lasso mean-squared error vs base-10 log of tuning parameter	19
13	Elastic net coefficients vs base-10 log of tuning parameter	20
14	Elastic-net mean-squared error vs base-10 log of tuning parameter	21
15	Group lasso coefficients vs base-10 log of tuning parameter	22
16	Mean-squared error vs base-10 log of tuning parameter	23

List of Tables

1	Constraint types in ALAMO	8
2	Description of features	10
3	Definition of feature groups in ALAMO	13
4	Defintion of group-constraints in ALAMO	14
5	Coefficients obtained from linear regression	15
6	Ridge coefficients as a function of tuning parameter	17
7	Lasso coefficients as a function of tuning parameter	19
8	Elastic net coefficients as a function of tuning parameter	21
9	Group-lasso coefficients as a function of tuning parameter	23
10	Solution paths of ALAMO	24

1 Abstract

This work presents the group-wise feature selection and constrained regression capabilities of ALAMO, an algebraic modeling framework developed at the Sahinidis Optimization group, Carnegie Mellon University and comparative benchmarks with popular feature selection techniques. Group-wise feature selection techniques are particularly useful for incorporating prior knowledge about the structure of features in terms of hierarchy and group-wise dependence in the model building process. The uniqueness of the approach is the superior control in terms of specifying groups of features and complex relationships between them, thereby allowing the modeler to define complex relationships among predictors and predictor groups. The algebraic model resulting from this approach contains fewer terms as compared to Generalized Linear Models (GLMs), thus minimizing generalization error and achieving higher model interpretability. Utilizing an iterative adaptive sampling approach optimized by internal derivative-free optimization solvers, this approach utilizes the subset-selection technique to find the feature subset that best fits the data. The constrained regression methodology in ALAMO is compared with popular modeling techniques such as linear regression, ridge, lasso, elastic-net and group lasso by fitting these models to the low birth weight dataset, a popular dataset for identifying maternal risk-factors affecting the weight of an infant. Basic feature transformations lead to group-wise structure among the features, thus necessitating a group-wise feature selection approach.

2 Introduction

Some of the major challenges in machine learning approaches include overfitting, high dimensionality and lack of interpretability. The aim of regularization is to reduce overfitting by either feature selection or shrinkage of predictor coefficients. Feature selection techniques also aim to incorporate model interpretability by setting the coefficients of certain predictors to zero, thereby excluding the entire term from the model. This technique, known as sparsity-based regularization, leads to faster computation, dimensionality reduction and improving model interpretability.

The first regularization technique dates back to 1970, when Hoerl et al.[1] included the ℓ_2 norm of the coefficients in the ordinary least-squares (OLS) regression problem to shrink the coefficients of the regression solution. This technique is also known as the ridge.

The Lasso[2] technique was invented in 1996 and became one of the highest cited papers in modern statistics and machine learning. Unlike the ridge which aims at shrinking individual coefficients in the regression problem, the lasso acts as a variable selection tool by including the ℓ_1 norm of the coefficient vector in the OLS objective function.

The elastic net, formulated in 2005 [3] aims at achieving a balance between the ridge and the lasso by including both, the ℓ_1 and ℓ_2 norms of the coefficient vector in the least-squares objective function. Thus, this is a tool that leads to The elastic net has quickly grown to become a popular tool in modeling financial modeling due to its unique functionality [4].

Yuan and Lin [5] invented the group-lasso technique for selection of feature groups and shrinkage of coefficients within selected feature groups. The model has been referenced in highly cited articles [6, 7] and has secured a unique position in the statistical machine learning community, owing to its fundamental nature and utility.

The best-subset selection[8, 9, 10] technique involves an exhaustive evaluation of all feature subsets to find the model that best fits the data according to a certain error metric. The best-subset selection has been widely regarded as the holy grail of statistics but is feared by computer scientists and modelers due to its intractable nature.

ALAMO[11, 12, 13] is an adaptive-sampling based algebraic modeling framework that utilizes derivative-free optimization techniques for minimizing various error criteria to come up with the simplest model.

The following sections describe each of these models in detail and experiments for constrained regression using ALAMO.

3 Regularization and feature selection

This section includes an overview of popular regularization methods and their relationship with feature selection and shrinkage.

3.1 The Ridge

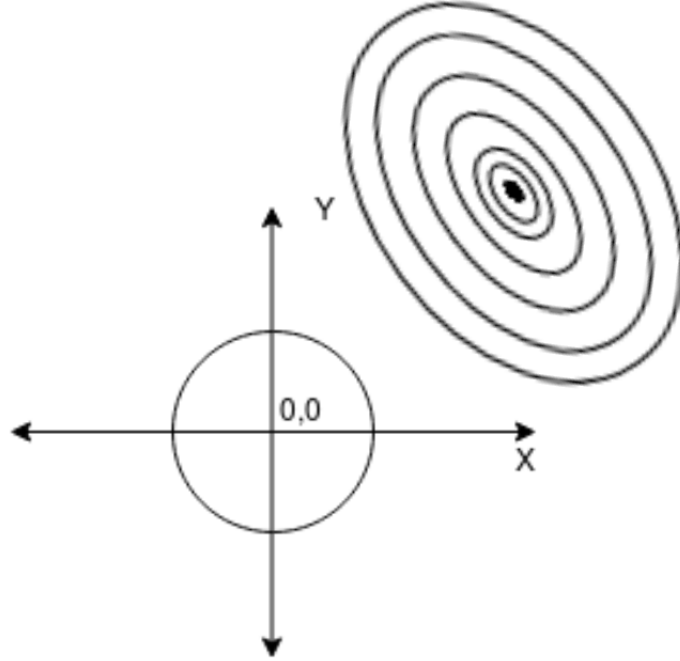


Figure 1: The ridge penalty as a constrained minimization problem

The ridge was formalized in 1970 by Hoerl et al.[1] and it seeks to minimize the following objective function:

$$\min_{\beta \in R^p} \left(\|\mathbf{y} - \beta_0 - \sum_{j=1}^J \mathbf{X}_j \beta_j\|_2^2 + \lambda \|\vec{\beta}\|_2^2 \right) \quad (1)$$

In the above equation, \mathbf{y} is the target variable vector, β_0 is the intercept, \mathbf{X} is the design matrix, J is the number of predictors, $\vec{\beta}$ is the coefficient vector and λ is the regularization parameter. The objective function includes the ℓ_2 norm of the coefficient vector and is responsible for shrinkage of coefficients. The degree of shrinkage is controlled by the regularization parameter λ . The ridge is a convex penalty and hence is tractable. The primary use of the ridge penalty is to shrink coefficients to reduce overfitting and thus minimize generalization error[14].

3.2 The Lasso

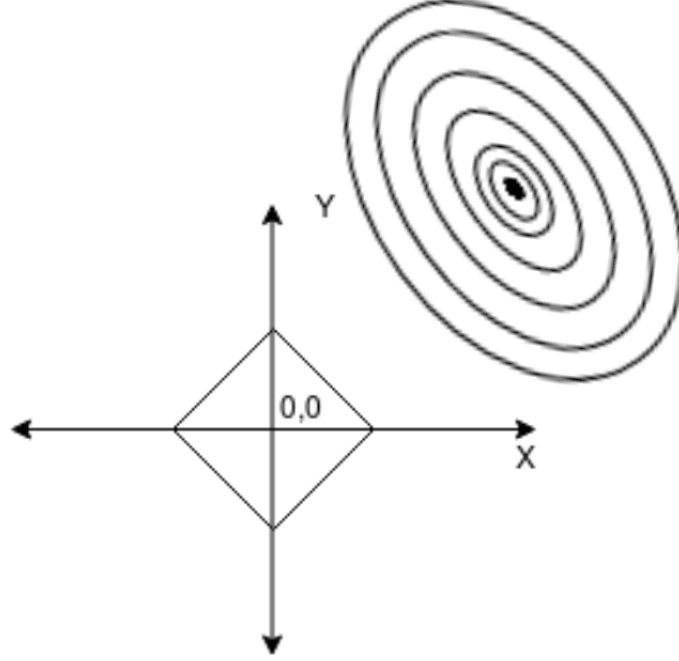


Figure 2: The lasso penalty as a constrained minimization problem

The lasso, formalized in 1996 by Tibshirani et al.[2] is a technique to select features by selectively setting predictor coefficients to zero. The formal definition of the lasso penalty is as follows:

$$\min_{\vec{\beta} \in R^p} \left(\|\mathbf{y} - \beta_0 - \sum_{j=1}^J \mathbf{X}_j \beta_j\|_2^2 + \lambda \|\vec{\beta}\|_1 \right) \quad (2)$$

Similar to the representation for the ridge, \mathbf{y} denotes the target variable vector, β_0 denotes the intercept, \mathbf{X} denotes the design matrix, J denotes the number of predictors, $\vec{\beta}$ denotes the coefficient vector and λ denotes the regularization parameter. Here, instead of the ℓ_2 norm, the objective function includes the ℓ_1 norm of the coefficient vector. This leads to setting individual coefficients to exactly zero along the lasso solution path and hence is used responsible for feature selection. The regularization parameter λ governs the degree to which coefficients are penalized; severe penalization leads to a very sparse model, leading to a high bias model, whereas very low penalization may lead to overfitting and a high generalization error. Like the ridge, the lasso is also a convex penalty and hence is tractable. The lasso has applications in feature selection and in reducing overfitting.

3.3 The Elastic Net

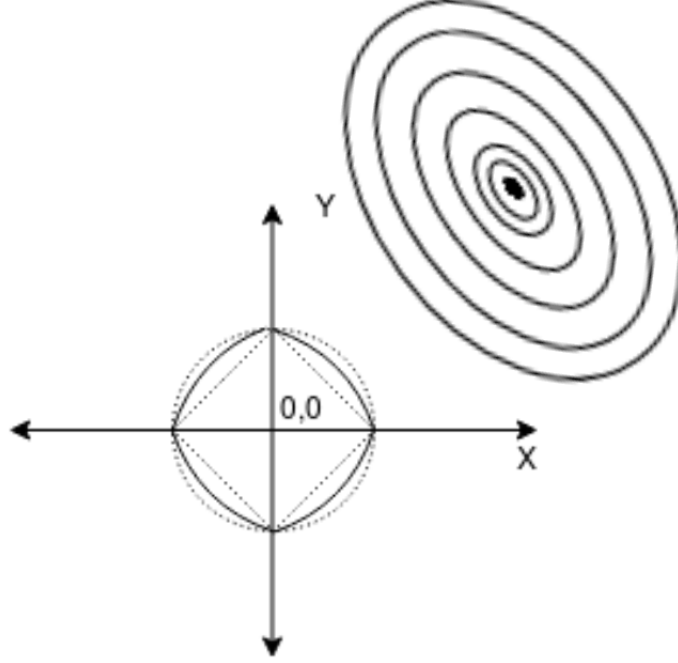


Figure 3: The elastic net penalty as a constrained minimization problem

The elastic net can be considered as an intermediate between the ridge and the lasso, addressing the shrinkage and variable selection problems simultaneously. The definition of the elastic net problem is as follows:

$$\min_{\vec{\beta} \in R^p} \left(\|\mathbf{y} - \beta_0 - \sum_{j=1}^J \mathbf{X}_j \beta_j\|_2^2 + \lambda_1 \|\vec{\beta}\|_1 + \lambda_2 \|\vec{\beta}\|_2^2 \right) \quad (3)$$

This representation is quite similar to those of the ridge and the lasso with the difference being that here two different penalty terms are included, instead of one. In the above equation, \mathbf{y} denotes the target variable vector, β_0 denotes the intercept, \mathbf{X} denotes the design matrix, J denotes the number of predictors, $\vec{\beta}$ denotes the coefficient vector and λ_1 denotes the regularization parameter for the ℓ_1 norm and λ_2 denotes the regularization parameter for the ℓ_2 norm. Alternatively, the two parameters can also be expressed as one parameter defining the ratio between ridge and lasso and another parameter defining the severity of the overall penalty. While the elastic net accounts for shrinkage and selection, it disregards the group-wise structure and relationships between predictors.

3.4 The Group-Lasso

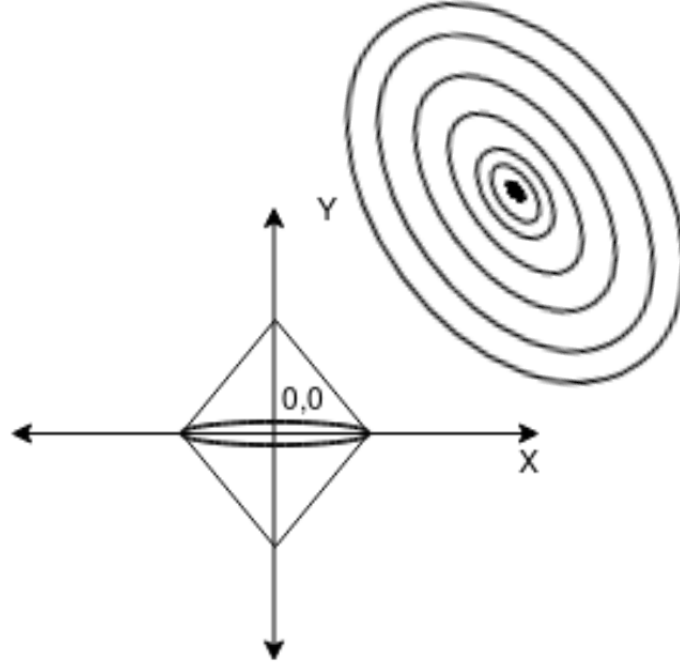


Figure 4: The group lasso penalty as a constrained minimization problem

The group lasso was formulated in 2006 by Yuan and Lin [5] and has become widely popular as a technique for selecting feature groups. The mathematical formulation of the group lasso is as follows:

$$\min_{\beta \in R^p} \left(\|\mathbf{y} - \beta_0 - \sum_{j=1}^J \mathbf{X}_j \beta_j\|_2^2 + \lambda \sum_{\ell=1}^L \sqrt{p_\ell} \|\vec{\beta}_\ell\|_2 \right) \quad (4)$$

In this representation, \mathbf{y} is the target variable, \mathbf{X} is the feature matrix, β_0 is the weight of the bias term and $\vec{\beta}_\ell$ is the subvector corresponding to the coefficients of variables occurring within the group ℓ . Here, J is the number of predictors and L is the number of predictor-groups. λ is the overall tuning parameter for the group lasso and $\sqrt{p_\ell}$ is the penalty proportional to the size (number of terms) in each group ℓ . This ensures normalizing the penalty according to the size of a group. The ℓ_2 penalty is applied within a group to reduce the magnitude of each of the weights within a group. These norms, being positive are directly added up and effectively work as an ℓ_1 norm across the groups of weights. This ℓ_1 leads to sparsity among feature groups. Thus, the group-lasso can be considered as a two-step penalty. The outer (ℓ_1) penalty is responsible for selecting feature groups and the inner (ℓ_2) penalty is responsible for coefficient shrinkage within the selected feature groups.

3.5 Best subset selection

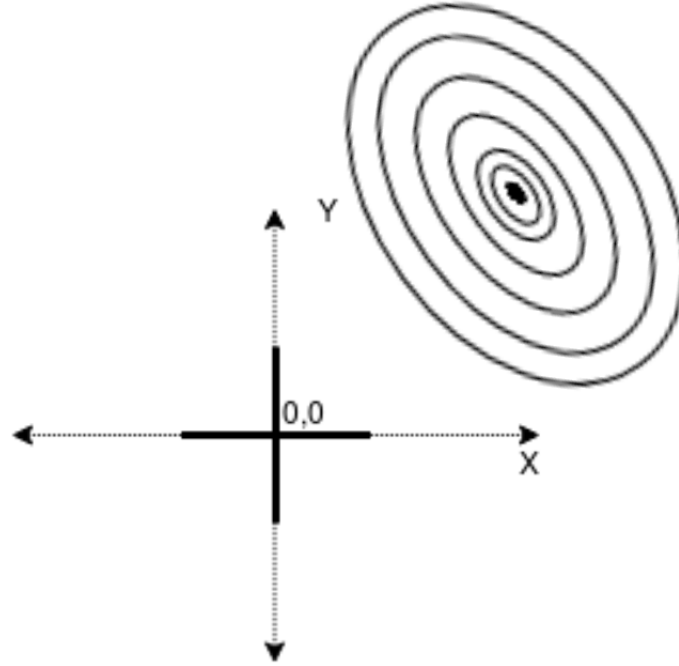


Figure 5: The best-subset selection technique

The best-subset technique [8, 9, 10] was first formalized in 1977 and as a technique for feature selection and is essentially an exhaustive search across all feature subsets to find the feature subset that minimizes the loss function with respect to the features vector consisting of the feature subset selected. While this isn't exactly the same as calculating a vector norm, it is known as the ℓ_0 norm owing to the effective feature set obtained. Each feature can either be included in the model or excluded from it. This leads to the following relation:

$$\text{Number of subsets} = 2^{\text{number of features}} \quad (5)$$

Being exponential in the number of features, the best subset selection technique is often considered intractable. However, due to the exhaustive search across feature subsets to find the best subset, this technique is widely considered as the holy grail of statistics and has been of interest lately due to advances in computing and enhancements in memory.

4 ALAMO (Automated Learning of Algebraic Models for Optimization)

ALAMO [11, 12, 13] is a black-box modeling toolbox developed at the Sahinidis Optimization Group, Carnegie Mellon University. ALAMO implements an integer-programming based best-subset selection strategy for variable selection which utilizes the hallowed ℓ_0 norm for variable selection and solves non-convex the optimization problem using derivative-free optimization solvers.

4.1 The ALAMO model-building process

The ALAMO model-building process consists of two key steps:

4.1.1 Surrogate Model Generation

The solution approach begins by sampling an initial set of points from the input data set, which is then used for building an initial model, starting with the lowest complexity. Various combinations of simple basis functions are considered for generating algebraic models, which are then solved using an optimization framework.

4.1.2 Adaptive Sampling

Adaptive sampling is an active learning [15, 16] approach which involves querying the data for selecting the next set of points such that maximum model accuracy can be obtained with minimum additional information. This is done by an error-maximization strategy, which means that the points which have the largest deviation from the model are included in the next sample for model building.

4.2 Constrained regression in ALAMO

ALAMO allows manual specification of constraints on features and includes specification of feature groups. These feature groups could be defined by either polynomial transformations of the same feature forming a group or based on prior knowledge about heirarchical structure within the features. The constraint types are as follows:

Keyword	Meaning
REQ	If group i is selected, group j is also selected.
XCL	If group i is selected, group j is NOT selected and vice-versa.
NMT	Not more than k variables within group i must be selected.
ATL	At least k variables within group i must be selected.

Table 1: Constraint types in ALAMO

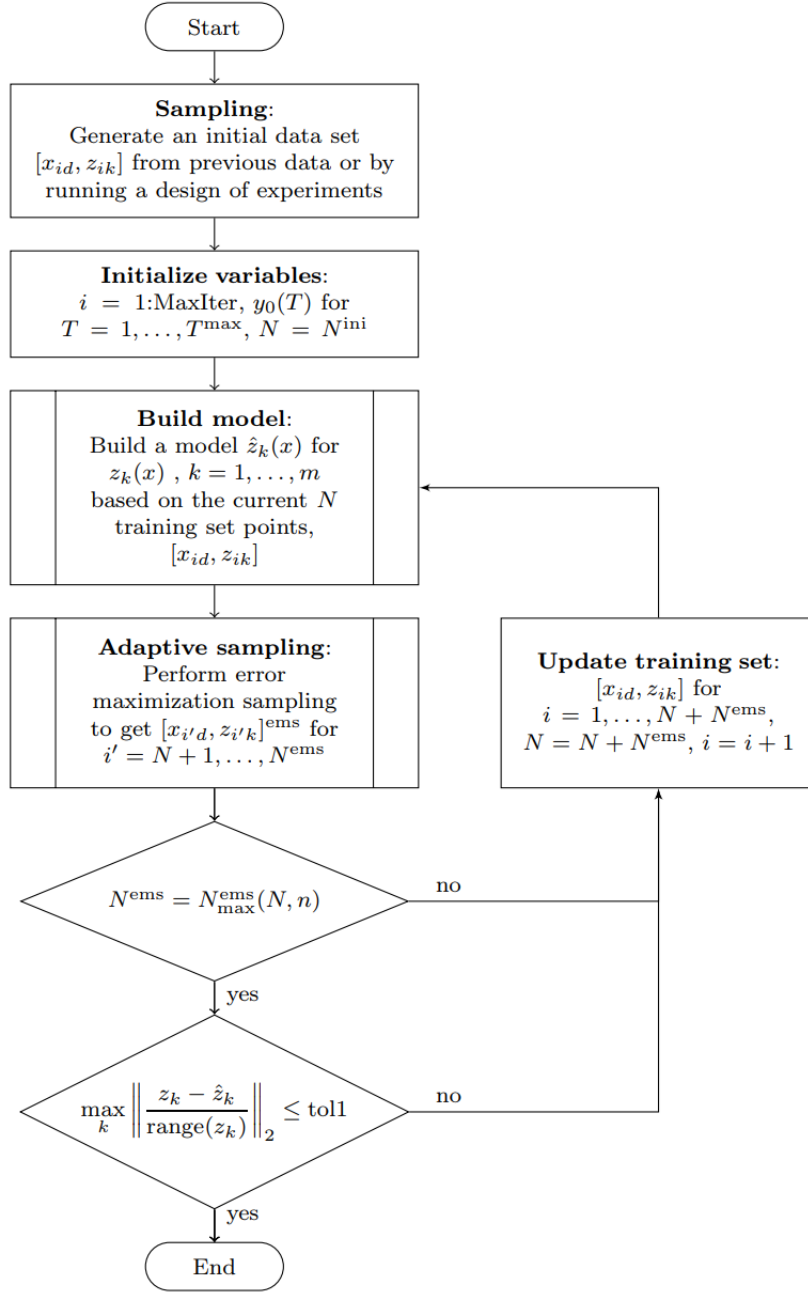


Figure 6: The ALAMO approach

5 Experiments

5.1 Dataset description

The dataset, popularly known as the low-birth-weight dataset consists of risk factors affecting the weight of a new born child. This dataset was first published in the book - Applied Logistic Regression by Hosmer and Lemeshow [17] and consists of the data were collected at Baystate Medical Center, Springfield, Mass during 1986. The data in the raw format can be found in popular repositories like MLData (<http://www.mldata.org/repository/data/viewslug/uci-20070111-lowbwt/>). The original dataset consists of a real valued target variable-weight of child, with 8 features and 189 data points. The features are as follows:

Feature	Abbreviation	Data type
Mothers age	age	continuous
Mothers weight	lwt	continuous
Smoking status during pregnancy	smoke	boolean
Mother's Race	black/white	categorical
Number of premature labors	ptl	integer
History of hypertension	ht	boolean
History of uterine irritability	ui	boolean
Number of clinician visits	ftv	integer

Table 2: Description of features

5.2 Feature transformations

Additional features corresponding to polynomial transformations of the features have been created. The features age1, age2, age3 correspond to first, second and third order polynomial transformation of the mother's age. Similarly, lwt2, lwt3 are features corresponding to square and cube of the mother's weight. The categorical features have been transformed using one-hot encoding. This transformed dataset is also available in the public domain for analysis.

5.2.1 Feature grouping

Feature groups are identified based on either polynomial transformations of the same primary variable or prior knowledge about association between features. For example, features age1, age2, age3 and features lwt1, lwt2, lwt3 form 2 groups based on polynomial transformations. Additionally, features corresponding to either one or more than one premature labors form a group and those corresponding to either one, two or more than two physician visits in the first trimester form another group. Lastly, one-hot encodings of a categorical variable form a group. The feature groups can be seen in the representation below:

	age			lwt			white	black	smoke	ptl1	ptl2m	ht	ui	ftv1	ftv2	ftv3m	bwt
0	-0.058334	0.011046	0.029562	0.124463	-0.021339	-0.130731	0	1	0	0	0	0	1	0	0	0	2.523
1	0.134366	0.055246	-0.096907	0.060067	-0.069228	-0.033348	0	0	0	0	0	0	0	0	0	1	2.551
2	-0.044570	-0.009415	0.045089	-0.059184	0.037463	0.004618	1	0	1	0	0	0	0	1	0	0	2.557
3	-0.030806	-0.026244	0.052490	-0.052029	0.023907	0.019035	1	0	1	0	0	0	1	0	1	0	2.594
4	-0.072099	0.035142	0.004822	-0.054414	0.028324	0.014572	1	0	1	0	0	0	1	0	0	0	2.600

Figure 7: Identification of feature groups

5.3 Correlation plot

Pearson correlation between each predictor and the target variable is plotted. The correlations can be seen as follows:

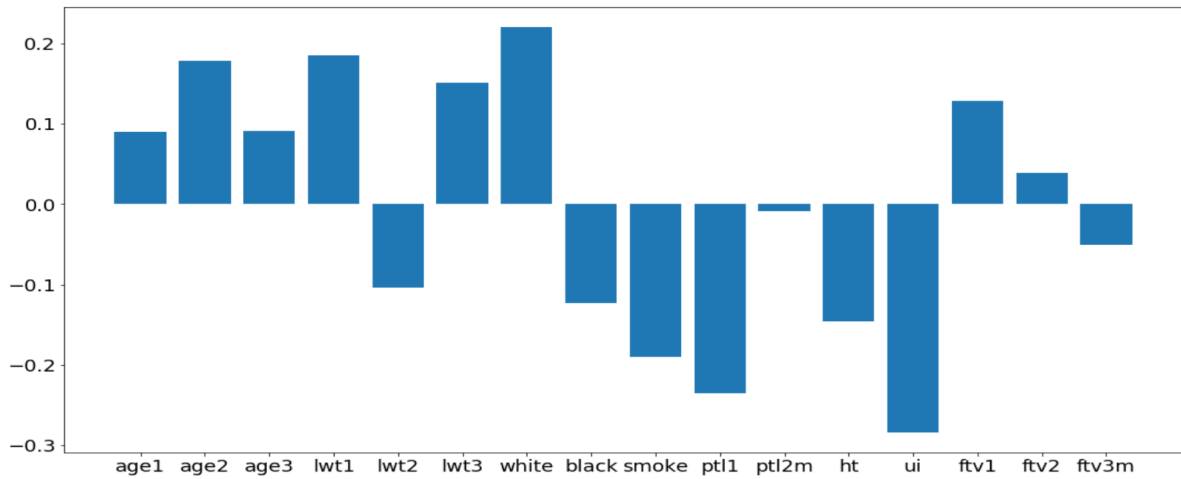


Figure 8: Correlation plot: Pearson correlation of each feature with target variable

It can be observed that the variable `ptl2m`, which corresponds to two or more premature labors is the variable least correlated with the response variable. Similarly, number of clinician visits have relatively low correlations with the response variable, weight. This is intuitive as we do not generally expect the number of times a mother visited the clinic to affect the weight of her new born child.

5.4 Modeling

Multiple models such as linear regression, ridge, lasso, elastic net, group-lasso and ALAMO were used for fitting the data and mean-squared error for each model was reported. Additionally, average runtime across 5 runs was measured for all models and reported. All experiments were run on a machine with Intel i7 CPU with 8 cores at 2.9 GHz and 16 GB RAM running 64-bit Ubuntu 18.04.

5.4.1 Generalized linear models

Generalized linear models such as multiple linear regression, ridge, lasso, elastic net and group-lasso were implemented. The data was shuffled, mean-normalized and a 4-fold cross validation mean-squared error was reported. For regularization models involving a parameter, a parametric sweep was conducted for various values of the tuning parameter and mean-squared errors were plotted. The minimum value of the mean-squared error from each model has been included in the model summary in a later section of this work. Linear regression, ridge, lasso and elastic net models were implemented using the Scikit-learn[18] framework in Python and the group lasso was implemented using the grpreg[19] package in R.

5.4.2 ALAMO

ALAMO version 2018.6.20 was licensed and downloaded from The Optimization Firm website: www.minlp.com.

Data and constraints were described in a .alm file. 16 groups were defined corresponding to each of the predictor variables and relationships between the predictors were defined to utilize the constrained regression capability in ALAMO. The group specification for ALAMO is different from that of the group lasso in that each variable is defined as a group entity and relationships between these groups are established. The groups are defined as follows:

Group-ID	Member-type	Member-indices
1	lin	1
2	lin	2
3	lin	3
4	lin	4
5	lin	5
6	lin	6
7	lin	7
8	lin	8
9	lin	9
10	lin	10
11	lin	11
12	lin	12
13	lin	13
14	lin	14
15	lin	15
16	lin	16

Table 3: Definition of feature groups in ALAMO

In the above representation the first column specifies the group ID. As each variable is defined as its own group, the group-ID is the same as variable-ID. The keyword lin specifies that only linear basis functions of the group are considered. Lastly, the column member-indices specifies the contents of each group in terms of the group-ID of each member. In this case, as each group contains only 1 variable and the member index is same as group-ID.

Group constraints can be specified between the predictors using pre-established relationships between predictors. The constraint definitions are as follows:

Group-ID	Output-ID	Constraint type	Integer parameter
1	1	atl	1
1	1	req	2
1	1	req	3
2	1	req	1
2	1	req	3
3	1	req	1
3	1	req	2
4	1	atl	1
4	1	req	5
4	1	req	6
5	1	req	4
5	1	req	6
6	1	req	4
6	1	req	5
7	1	atl	1
7	1	xcl	8
8	1	xcl	7
10	1	req	11
11	1	req	10
14	1	req	15
15	1	req	16
16	1	req	14

Table 4: Defintion of group-constraints in ALAMO

In this representation, group-ID represents the group-ID defined in the group definitions section, output-ID specifies the number of outputs for which the constraint would be imposed. This specification is particularly useful in multivariate linear regression. The constraint-type section defines the type of group-constraint to be utilized. A summary of group constraints can be seen in section 4.3 of this work. Lastly, the integer-parameter column contains the index of the constraint to which the group belongs. As each group could belong to multiple constraints, this column can contain multiple group-ids. In our case, as each inter-group relation is symmetric and transitive, these constraints can also be defined in a cyclic manner: $(1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 1)$.

6 Results and Discussion

6.1 Linear Regression

Linear regression can be considered as the simplest of all the models considered in this work. In this model, a multiple linear regression model is fitted using the data. The coefficients obtained are as follows:

Feature	Weight
age1	-0.4211
age2	1.7217
age3	0.7250
lwt1	2.3606
lwt2	0.2525
lwt3	1.4769
white	0.1749
black	-0.2669
smoke	-0.2165
ptl1	-0.4291
ptl2m	0.2897
ht	-0.6244
ui	-0.5241
ftv1	0.2067
ftv2	-0.0392
ftv3m	-0.1249
intercept	3.1077
mse	0.4825

Table 5: Coefficients obtained from linear regression

The simplest form of multiple linear regression contains non-zero values for all predictor coefficients, indicating that all terms are included in the model. The 4-fold cross-validation mean-squared error is obtained by as 0.4825. The average overall runtime for linear regression is 5 ms.

As this model doesn't include any regularization, it tends to overfit on the training data and hence gives a relatively high generalization error. This model can be improved upon by using regularization methods and incorporating group-wise structure among features to build a better model.

6.2 Ridge

The ridge is a convex penalty responsible for shrinkage of coefficients. The magnitude of the coefficients is determined by λ , the penalizing coefficient. In the experiment, a parametric sweep across different penalizing coefficient values starting at 0.001 and increasing in steps of multiples of 10 to a value of $\lambda = 1000$. Average value of coefficients obtained for all predictors from a four-fold cross-validation are plotted against the base-10 log of the penalizing coefficient value. This gives the following graph:

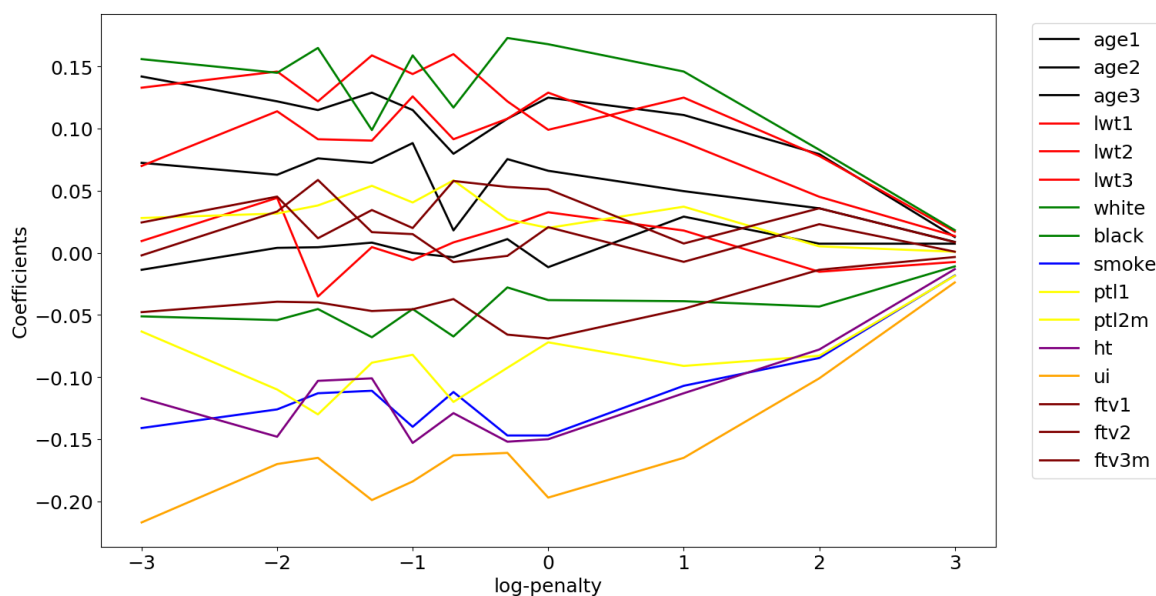


Figure 9: Ridge coefficients vs base-10 log of tuning parameter

In the above graph, the predictors are color-coded according to the groups they belong to. Although it does not affect the ridge penalty this representation is useful in comparing the ridge with group-wise feature selection discussed in a later section. It can be seen that the ridge doesn't set any coefficient to exactly zero, however, as the penalizing coefficient value increases, it can be seen that the magnitude of the coefficients decreases. The value of each coefficient as it varies with the tuning parameter is described in the following table:

penalty	0.001	0.01	0.02	0.05	0.10	0.20	0.50	1.00	10.00	100.00	1,000.00
log-penalty	-3.0000	-2.0000	-1.6990	-1.3010	-1.0000	-0.6990	-0.3010	0.0000	1.0000	2.0000	3.0000
age1	-0.0136	0.0041	0.0046	0.0083	0.0000	-0.0035	0.0111	-0.0115	0.0292	0.0074	0.0074
age2	0.1416	0.1218	0.1147	0.1286	0.1149	0.0798	0.1083	0.1246	0.1111	0.0795	0.0128
age3	0.0725	0.0629	0.0761	0.0725	0.0884	0.0181	0.0755	0.0661	0.0497	0.0359	0.0088
lwt1	0.1327	0.1459	0.1222	0.1587	0.1438	0.1604	0.1223	0.0991	0.1246	0.0780	0.0171
lwt2	0.0095	0.0443	-0.0351	0.0047	-0.0058	0.0085	0.0212	0.0327	0.0180	-0.0152	-0.0073
lwt3	0.0700	0.1139	0.0915	0.0904	0.1262	0.0915	0.1083	0.1287	0.0892	0.0452	0.0136
white	0.1555	0.1448	0.1653	0.0989	0.1585	0.1165	0.1730	0.1680	0.1456	0.0830	0.0183
black	-0.0511	-0.0541	-0.0452	-0.0679	-0.0452	-0.0673	-0.0278	-0.0380	-0.0389	-0.0432	-0.0109
smoke	-0.1406	-0.1262	-0.1128	-0.1111	-0.1400	-0.1119	-0.1469	-0.1472	-0.1072	-0.0846	-0.0182
ptl1	-0.0633	-0.1102	-0.1301	-0.0884	-0.0820	-0.1196	-0.0925	-0.0720	-0.0910	-0.0828	-0.0185
ptl2m	0.0281	0.0317	0.0382	0.0540	0.0406	0.0584	0.0270	0.0201	0.0372	0.0053	0.0009
ht	-0.1169	-0.1480	-0.1028	-0.1014	-0.1533	-0.1288	-0.1522	-0.1500	-0.1133	-0.0778	-0.0131
ui	-0.2172	-0.1705	-0.1654	-0.1995	-0.1844	-0.1633	-0.1606	-0.1965	-0.1647	-0.1008	-0.0238
ftv1	0.0245	0.0453	0.0118	0.0345	0.0200	0.0579	0.0531	0.0512	0.0075	0.0359	0.0086
ftv2	-0.0020	0.0333	0.0587	0.0167	0.0151	-0.0074	-0.0024	0.0207	-0.0072	0.0231	0.0010
ftv3m	-0.0477	-0.0393	-0.0399	-0.0468	-0.0454	-0.0372	-0.0658	-0.0689	-0.0450	-0.0136	-0.0033
intercept	2.9643	2.9504	2.9594	2.9589	2.9569	2.9493	2.9661	2.9551	2.9340	2.9552	2.9537
mse	0.4543	0.4687	0.5274	0.4634	0.4825	0.4666	0.4725	0.5086	0.4676	0.4235	0.4858

Table 6: Ridge coefficients as a function of tuning parameter

We can also plot the mean squared error as a function of the penalizing coefficient λ . This helps us obtain the following graph:

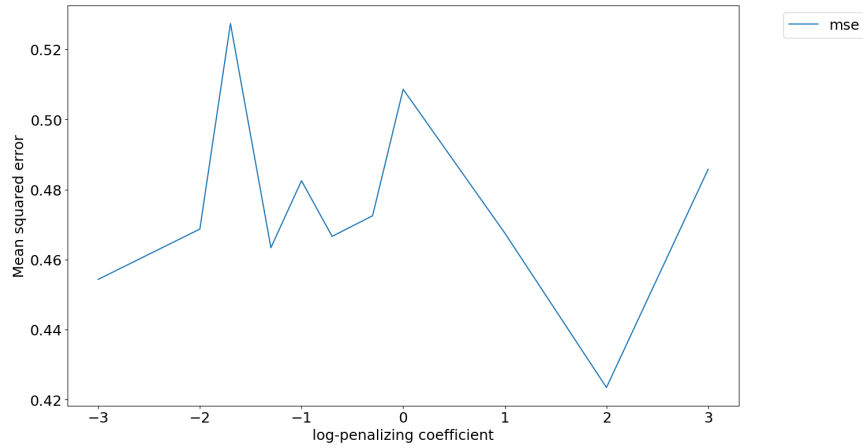


Figure 10: Ridge mean-squared error vs base-10 log of tuning parameter

We can infer that the lowest value of mean-squared error is obtained when $\lambda = 100$ and the mean squared error value is 0.4235. This mean squared error is lower than that of linear regression as coefficient shrinkage helps in minimizing generalization error. The average runtime for the ridge is 43.1 ms.

6.3 Lasso

The lasso is responsible for feature selection by setting the coefficients of predictors in the model to exactly zero, thereby introducing sparsity in the model. This is evident from the results and can be seen by plotting the value of all coefficients as a function of the log base-10 penalizing coefficient.

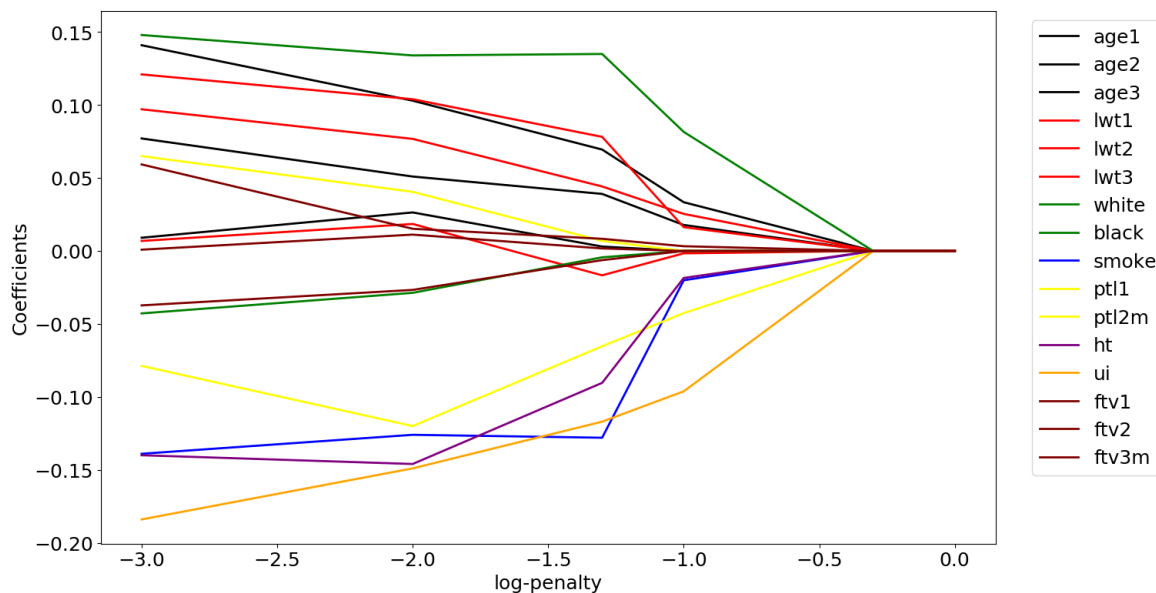


Figure 11: Lasso coefficients vs base-10 log of tuning parameter

In the above graph, individual predictors are color-coded according to the group they belong to. From the figure, we can analyze how different coefficients are set to zero as the penalizing coefficient λ is increased from 0.0001 to 1 in multiples of 10. The sparsity introduced in the model can further be explained by the following table indicating the coefficients, intercept and mean-squared error as a function of the tuning parameter λ .

penalty	0.0001	0.0010	0.0100	0.0500	0.1000	0.5000	1.0000
log-penalty	-4.0000	-3.0000	-2.0000	-1.3010	-1.0000	-0.3010	0.0000
age1	0.0430	0.0066	-0.0126	0.0000	0.0000	0.0000	0.0000
age2	0.1535	0.1303	0.0945	0.0376	0.0317	0.0000	0.0000
age3	0.1044	0.0599	0.0541	0.0188	0.0014	0.0000	0.0000
lwt1	0.1447	0.1352	0.1267	0.0606	0.0364	0.0000	0.0000
lwt2	-0.0254	-0.0160	0.0014	0.0000	-0.0044	0.0000	0.0000
lwt3	0.1280	0.1223	0.0901	0.0416	0.0062	0.0000	0.0000
white	0.1337	0.1363	0.1433	0.0826	0.0582	0.0000	0.0000
black	-0.0769	-0.0451	-0.0509	-0.0132	0.0000	0.0000	0.0000
smoke	-0.1508	-0.1150	-0.1386	-0.0867	-0.0468	0.0000	0.0000
ptl1	-0.0744	-0.0958	-0.1088	-0.0626	-0.0279	0.0000	0.0000
ptl2m	0.0625	0.0112	0.0150	0.0000	0.0000	0.0000	0.0000
ht	-0.1206	-0.1322	-0.1371	-0.1021	-0.0177	0.0000	0.0000
ui	-0.1573	-0.1916	-0.1545	-0.1423	-0.0911	0.0000	0.0000
ftv1	0.0359	0.0310	0.0390	0.0124	0.0027	0.0000	0.0000
ftv2	0.0442	0.0082	0.0220	0.0000	0.0000	0.0000	0.0000
ftv3m	-0.0557	-0.0337	-0.0115	-0.0082	0.0000	0.0000	0.0000
intercept	2.9483	2.9403	2.9324	2.9327	2.9535	2.9650	2.9613
mse	0.4942	0.4181	0.5108	0.5109	0.4894	0.5483	0.4417

Table 7: Lasso coefficients as a function of tuning parameter

We can also plot the mean-square error as a function of the base-10 log of the tuning parameter . The graph obtained is as follows:

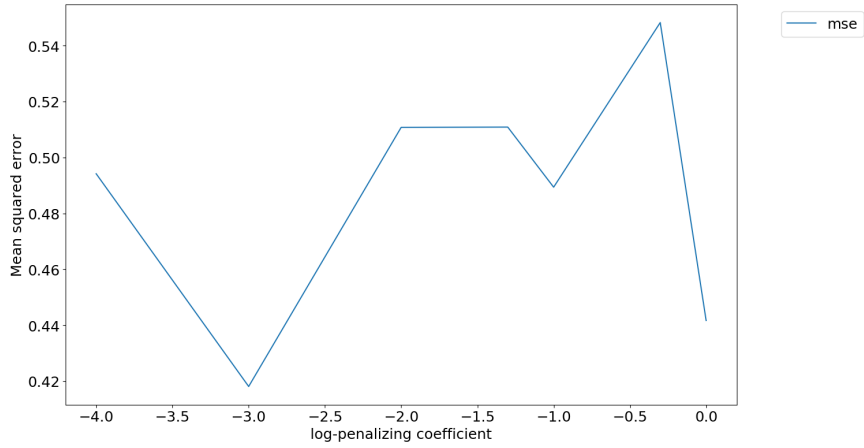


Figure 12: Lasso mean-squared error vs base-10 log of tuning parameter

The above table indicates that the minimum mean-squared error is obtained when the penalizing coefficient $\lambda = 0.001$ and the value of the mean squared error is 0.4181. The average overall runtime of the lasso is 28.4 ms.

6.4 Elastic Net

The elastic net utilizes a penalty that involves both, the ℓ_1 and ℓ_2 norms, thereby allowing for a better tunability of the model. In this implementation of the elastic net, the ratio between the ℓ_1 and ℓ_2 norms is set at 1:1 but can also be tuned as an additional hyper-parameter.

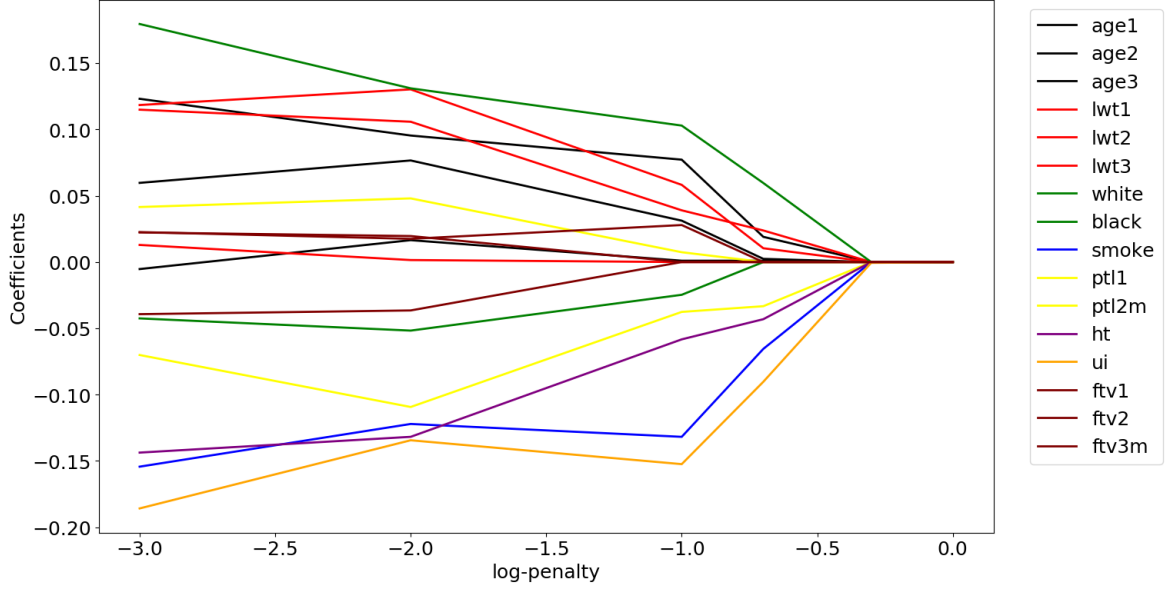


Figure 13: Elastic net coefficients vs base-10 log of tuning parameter

The above figure indicates the sparsity introduced in the model as a function of the base-10 log of the penalizing coefficient. As the penalizing coefficient λ increases, more and more coefficients are set to 0. Along with feature selection, the elastic net also penalizes the selected features, thereby leading to shrinkage. This leads to a model that generalizes well.

However, the elastic net penalty, by nature, does not account for group-wise relationships between variables. As a result, the variables elimination process is independent of the group-wise structure of features, leading to unstructured sparsity. The sparsity pattern of the elastic net can be observed in the following table:

penalty	0.0010	0.0100	0.1000	0.2000	0.5000	1.0000
log-penalty	-3.0000	-2.0000	-1.0000	-0.6990	-0.3010	0.0000
age1	-0.0053	0.0165	0.0010	0.0010	0.0000	0.0000
age2	0.1231	0.0954	0.0772	0.0190	0.0000	0.0000
age3	0.0597	0.0766	0.0312	0.0024	0.0000	0.0000
lwt1	0.1184	0.1302	0.0582	0.0104	0.0000	0.0000
lwt2	0.0129	0.0015	0.0000	0.0000	0.0000	0.0000
lwt3	0.1149	0.1058	0.0390	0.0238	0.0000	0.0000
white	0.1795	0.1310	0.1029	0.0596	0.0000	0.0000
black	-0.0425	-0.0517	-0.0247	0.0000	0.0000	0.0000
smoke	-0.1543	-0.1221	-0.1318	-0.0655	0.0000	0.0000
ptl1	-0.0701	-0.1093	-0.0376	-0.0333	0.0000	0.0000
ptl2m	0.0415	0.0480	0.0074	0.0000	0.0000	0.0000
ht	-0.1437	-0.1318	-0.0583	-0.0431	0.0000	0.0000
ui	-0.1858	-0.1344	-0.1524	-0.0903	-0.0006	0.0000
ftv1	0.0226	0.0175	0.0279	0.0000	0.0000	0.0000
ftv2	0.0223	0.0196	0.0000	0.0000	0.0000	0.0000
ftv3m	-0.0393	-0.0365	0.0000	0.0000	0.0000	0.0000
intercept	2.9584	2.9300	2.9424	2.9418	2.9271	2.9289
mse	0.4623	0.4191	0.4018	0.4776	0.5481	0.4153

Table 8: Elastic net coefficients as a function of tuning parameter

We can also plot the mean-squared error as a function of the tuning parameter to obtain the following graph:

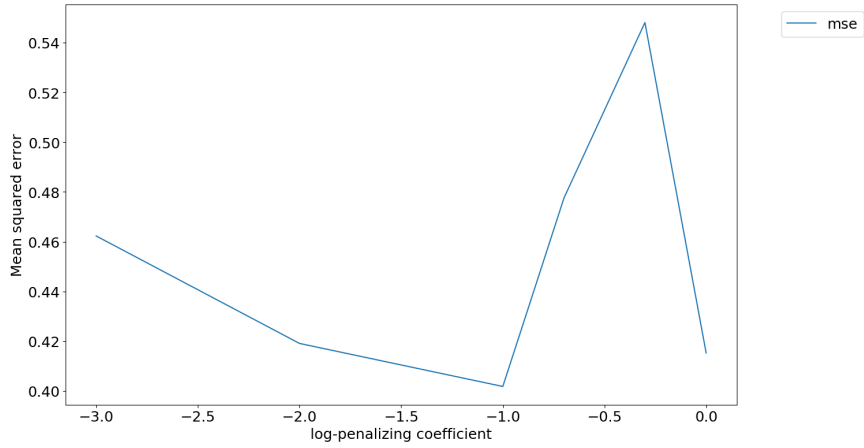


Figure 14: Elastic-net mean-squared error vs base-10 log of tuning parameter

It can be seen that the minimum error is achieved at $\lambda = 0.1$ and the error value is 0.4018. The average overall runtime of the elastic net is found to be 25.8 ms.

6.5 Group Lasso

The group-lasso allows for specification of feature groups in the penalty, allowing for a model that is interpretable and accounts for relationships between features in the model building process. The coefficients values for all predictors are plotted below as the tuning parameter is varied from 0.001 to 1 in steps of multiples of 10.

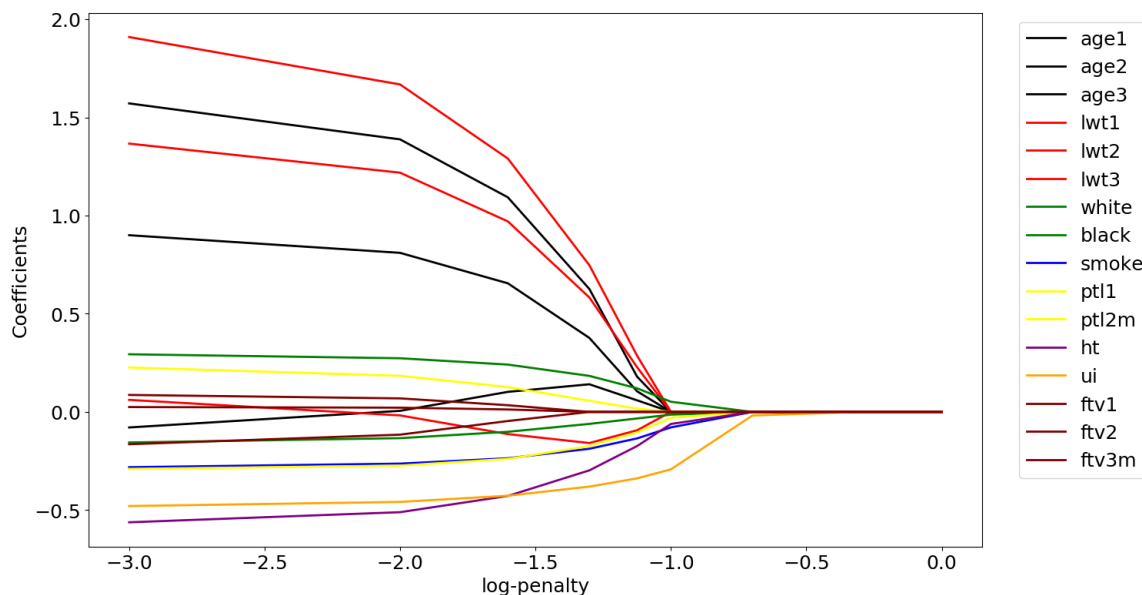


Figure 15: Group lasso coefficients vs base-10 log of tuning parameter

The most striking difference between the group lasso and other models discussed earlier is in the way in which the predictors are eliminated from the model. We can see from the above graph that coefficients of predictors belonging to a group are simultaneously set to zero as the value of the tuning parameter increases. For example, we can see in the coefficient table below that the first group of variables eliminated from the model is the group corresponding to number of clinician visits. We know that if a visit to the physician does not affect the weight of the infant, two visits or more should also not affect the weight of the infant. While all models discussed earlier fail to capture this insight, the group lasso sets the coefficients corresponding to ftv1, ftv2, ftv3m simultaneously to zero, producing a model that is easier to interpret.

penalty	0.0010	0.0100	0.0250	0.0500	0.0750	0.1000	0.2000	0.5000	1.0000
log-penalty	-3.0000	-2.0000	-1.6021	-1.3010	-1.1249	-1.0000	-0.6990	-0.3010	0.0000
intercept	3.0487	3.0434	3.0379	3.0289	3.0149	3.0021	2.9473	2.9446	2.9446
age1	-0.0792	0.0054	0.1020	0.1407	0.0590	0.0000	0.0000	0.0000	0.0000
age2	1.5714	1.3883	1.0931	0.6260	0.1796	0.0000	0.0000	0.0000	0.0000
age3	0.9001	0.8102	0.6551	0.3767	0.1049	0.0000	0.0000	0.0000	0.0000
lwt1	1.9094	1.6682	1.2907	0.7470	0.2866	0.0000	0.0000	0.0000	0.0000
lwt2	0.0614	-0.0179	-0.1129	-0.1585	-0.0935	0.0000	0.0000	0.0000	0.0000
lwt3	1.3667	1.2186	0.9706	0.5829	0.2278	0.0000	0.0000	0.0000	0.0000
white	0.2935	0.2733	0.2411	0.1835	0.1203	0.0517	0.0000	0.0000	0.0000
black	-0.1556	-0.1337	-0.1013	-0.0610	-0.0334	-0.0140	0.0000	0.0000	0.0000
smoke	-0.2816	-0.2635	-0.2361	-0.1878	-0.1350	-0.0788	0.0000	0.0000	0.0000
ptl1	-0.2903	-0.2736	-0.2394	-0.1743	-0.1045	-0.0301	0.0000	0.0000	0.0000
ptl2m	0.2261	0.1840	0.1267	0.0570	0.0164	0.0014	0.0000	0.0000	0.0000
ht	-0.5623	-0.5109	-0.4281	-0.2977	-0.1740	-0.0614	0.0000	0.0000	0.0000
ui	-0.4795	-0.4586	-0.4270	-0.3805	-0.3387	-0.2925	-0.0183	0.0000	0.0000
ftv1	0.0864	0.0688	0.0340	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ftv2	0.0247	0.0215	0.0124	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ftv3m	-0.1647	-0.1159	-0.0465	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
mse	0.4332	0.4196	0.4377	0.4560	0.4713	0.4961	0.5299	0.5329	0.5329

Table 9: Group-lasso coefficients as a function of tuning parameter

We can also plot the mean-squared error as a function of the base-10 log of the tuning parameter to obtain the following graph:

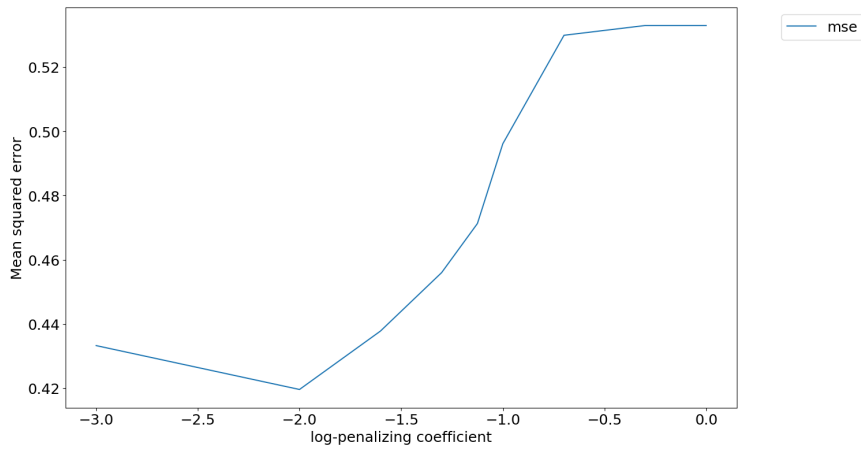


Figure 16: Mean-squared error vs base-10 log of tuning parameter

We can see that the minimum value for mean-squared error is obtained when the tuning parameter $\lambda = 0.01$ and the value of the error is 0.4196. This is slightly higher than the error value obtained from the elastic net. The average overall runtime of the group-lasso is 74.0 ms.

6.6 ALAMO

We can analyze which features are included in the model as a function of each ALAMO iteration. The following table shows the feature subsets and mean-squared error as ALAMO iterations progress:

Iteration #	Feature subsets	MSE
1	(age1, age2, age3, lwt1, lwt2, lwt3, white)	4.37
2	(age1, age2, age3, lwt1, lwt2, lwt3, white, 1)	0.472
3	(age1, age2, age3, lwt1, lwt2, lwt3, white, ui, 1)	0.444
4	(age1, age2, age3, lwt1, lwt2, lwt3, white, ht, ui, 1)	0.420
5	(age1, age2, age3, lwt1, lwt2, lwt3, white, smoke, ht, ui, 1)	0.402

Table 10: Solution paths of ALAMO

The linear model obtained from ALAMO is as follows:

$$Y(\text{predicted}) = -0.23 * \text{age1} + 1.5 * \text{age2} + 1.1 * \text{age3} + 1.7 * \text{lwt1} + 0.23 * \text{lwt2} + 1.3 * \text{lwt3} + 0.40 * \text{white} - 0.36 * \text{smoke} - 0.60 * \text{ht} - 0.50 * \text{ui} + 3.0$$

We can observe that ALAMO renders the model with the fewest terms as compared with the previous models, while maintaining group-wise relationships between predictors. It also achieves a mean-squared error lower than the group-lasso and comparable with the elastic-net. ALAMO achieves structured sparsity among predictors as apposed to unstructured sparsity with the elastic net. The overall result is a model that contains the fewest terms while maintaining a low generalization error and accounts for constraints among predictors and predictor-groups.

While ALAMO converges in fewer iterations as compared to other algorithms, each ALAMO iteration is computationally expensive, owing to the non-convexity of the subset selection problem. As a result, the average overall runtime of ALAMO is 334.0 ms and is slower than the models described earlier.

7 Conclusion

The following table summarizes the results from the multiple models used for fitting the data:

Model	Mean-squared error	Average Runtime (ms)	Number of terms
Linear regression	0.4825	5.9	16
Ridge	0.4235	43.1	16
Lasso	0.4181	28.4	16
Elastic net	0.4018	25.8	13
Group lasso	0.4196	74.0	16
ALAMO	0.4020	334.0	11

We can see that linear regression contains 16 terms and has the highest generalization error due to overfitting on the training data. However, being the simplest model to train, linear regression is the fastest and has an average total runtime of 5.9 ms.

The ridge attempts to fix the problem of overfitting by shrinking the coefficients and this can be seen by the lower mean-squared error of the ridge over the simple linear regression model. The ridge does not remove predictors from the model and hence the number of terms is still 16. As it solves the additional convex penalty term, the ridge takes longer to run than the linear regression model and this is evident from the runtime results.

The lasso gives a mean-squared error lower than the ridge and linear regression when the number of terms included is 16. However, as the model gets sparser, the error goes on increasing due to underfitting. Here, the trade-off between model interpretability and model accuracy comes into picture. Intuitively, the terms corresponding to the number of clinician visits should be excluded from the model, but doing so leads to an increase in mean-squared error if the lasso is used for feature selection.

The elastic net serves as an intermediate penalty between the lasso and the ridge and accounts for both, variable shrinkage and selection. It gives a lower mean-squared error than lasso, ridge and linear regression as it allows for better tunability and incorporates shrinkage and selection. It leads to a model with 13 terms. However, the elastic net and the lasso do not account for incorporation of prior knowledge about the group-wise nature of predictors in the model building process.

The group lasso accounts for the group-wise relationship between the predictors. The lowest mean-squared error from the group-lasso is higher than that of the elastic net due to the constraint that individual variables cannot be removed from the model, thus leading to a slightly higher error at the advantage of better model interpretability. As the penalty is increased, the number of terms in the group-lasso decreases but the error increases due to underfitting. The runtime of the group lasso is also higher than that of other GLMs due to the complexity of the penalty term.

The best-subset strategy of ALAMO gives a mean-squared error of 0.402 and leads to a model with 11 terms. Owing to manual specification of constraints on variables and variable groups, ALAMO leads to a model that is more interpretable and allows for better control by incorporating prior knowledge and insight about features and feature relationships. Owing to the non-convex penalty that ALAMO tries to solve, the runtime of ALAMO is greater than GLMs discussed earlier.

All models discussed above have their unique characteristics. The group lasso and ALAMO are two models that allow for defining feature relationships that can be utilized in the model building process. The group lasso incorporates structured sparsity-based regularization and allows for variable and group specification. However, the group selection is entirely dependent on a single tuning parameter and poses limitations in terms of defining complex relationships between predictors. ALAMO, on the other hand, allows for specification of group-constraints at a predictor-level and also accounts for constraints on the presence and absence of certain groups of predictors in the model. This leads to a lower error and higher control on the model in comparison with the group-lasso at the cost of higher runtime. Additionally, ALAMO allows for non-linear transformations of the predictors, allowing for learning more complex functions from data. However, this analysis is beyond the scope of this work.

References

- [1] Arthur E Hoerl and Robert W Kennard. “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1 (1970), pp. 55–67.
- [2] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.
- [3] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.
- [4] Hyun Hak Kim and Norman R Swanson. “Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence”. In: *Journal of Econometrics* 178 (2014), pp. 352–367.
- [5] Ming Yuan and Yi Lin. “Model selection and estimation in regression with grouped variables”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67.
- [6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. “Unsupervised learning”. In: *The elements of statistical learning*. Springer, 2009, pp. 485–585.
- [7] Stephen Boyd et al. “Distributed optimization and statistical learning via the alternating direction method of multipliers”. In: *Foundations and Trends® in Machine learning* 3.1 (2011), pp. 1–122.
- [8] Patrenahalli M. Narendra and Keinosuke Fukunaga. “A branch and bound algorithm for feature subset selection”. In: *IEEE Transactions on computers* 9 (1977), pp. 917–922.
- [9] George H John, Ron Kohavi, and Karl Pfleger. “Irrelevant features and the subset selection problem”. In: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 121–129.
- [10] Ron Kohavi and George H John. “Wrappers for feature subset selection”. In: *Artificial intelligence* 97.1-2 (1997), pp. 273–324.
- [11] Alison Cozad, Nikolaos V Sahinidis, and David C Miller. “Learning surrogate models for simulation-based optimization”. In: *AIChE Journal* 60.6 (2014), pp. 2211–2227.
- [12] Alison Cozad, Nikolaos V Sahinidis, and David C Miller. “A combined first-principles and data-driven approach to model building”. In: *Computers & Chemical Engineering* 73 (2015), pp. 116–127.
- [13] Zachary T Wilson and Nikolaos V Sahinidis. “The ALAMO approach to machine learning”. In: *Computers & Chemical Engineering* 106 (2017), pp. 785–795.
- [14] Craig Saunders, Alexander Gammerman, and Volodya Vovk. “Ridge regression learning algorithm in dual variables”. In: (1998).
- [15] Dana Angluin. “Queries and concept learning”. In: *Machine learning* 2.4 (1988), pp. 319–342.

- [16] Burr Settles. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 2009.
- [17] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.
- [18] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [19] Patrick Breheny and Jian Huang. “Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors”. In: *Statistics and computing* 25.2 (2015), pp. 173–187.