A Comparison of different Ridge parameters in an Asthma Persistence Prediction model

Ioannis I. Spyroglou, Eleni A. Chatzimichail, E.N. Paraskakis, and Alexandros G. Rigas

***Abstract*—**The purpose of this paper is to investigate the important role of the ridge parameter in a logistic regression model by comparing several different ridge parameters. These are applied in the study of an asthma persistent prediction problem. High collinearity among the explanatory variables leads to the use of a logistic ridge regression model in order to obtain better predictions. The use of different ridge parameters results to different logistic ridge regression models which predict asthma with different accuracies, as far as positive and negative predictive values are concerned. Additionally, the most interesting conclusion in using different ridge parameters for constructing the logistic ridge regression model, is the existence of different factors which are statistically significant, making the asthma persistence prediction problem more complex. For the evaluation of the model, a method which combines bootstrapping and randomized quantile residuals of the estimated models is used.

***Keywords*—**Asthma outcome, Multicollinearity, Logistic Ridge regression, Ridge Parameter, Randomized Quantile Residuals, Bootstrap

I. Introduction

Multicollinearity is one of the most important matters when the number of the explanatory variables is large and the correlations between them are strong and significant. In order to deal with multicollinearity introduced in 1934 by Frisch [1] a ridge regression model may be used. Ridge regression [2-3] is a shrinkage technique for analyzing data that suffer from significant collinearity between the predictor variables that makes the maximum likelihood approach unstable because the standard errors of the estimated coefficients become very large.

The most difficult task in Ridge regression is to determine the ridge parameter. Hoerl and Kennard proved that when collinearity exists there is always a model for ridge parameter $\lambda>0$ for which the MSE is less than the MSE of the unrestricted model [2-3]. Many articles have been proposing different estimates for the ridge parameter [4-13]. Although there are many proposals for using a ridge parameter, most of them are used for linear regression models.

In [14] the authors apply the ridge parameters proposed in [2] and [15] in logistic regression. Also in [16] a number of logistic ridge regression parameters are applied and investigated through a Monte Carlo simulation. In addition the ridge parameter can be estimated by using a cross – validation technique for the calculation of the minimum mean squared error ($\mathrm{MSE_{cv}}$), the mean minus log-likelihood ($\mathrm{MML_{cv}}$) and the mean classification error ($\mathrm{MCE_{cv}}$) [17].

The variance of the estimated regression coefficients is calculated through a bootstrap method [18,26,32] which uses the randomized quantile residuals. This is necessary because there is extra uncertainty due to the large number of explanatory variables. The randomized quantile residuals follow the standard normal distribution and are useful in testing the validity of the model [19].

Ioannis Spyroglou is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (corresponding author to provide phone: +306955954849 ; e-mail: ispyrogl@ee.duth.gr).

Eleni Chatzimichail is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (e-mail: echatzim@ee.duth.gr).

Emmanouil N. Paraskakis, Prof., is with the Medical School , Democritus University of Thrace, Alexandroupolis, CO 68100 GRRECE (e-mail: eparaska@med.duth.gr).

Alexandros Rigas, Prof., is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (e-mail: rigas@ee.duth.gr).

II. Materials And Methods

A. Clinical Data

Data from 148 patients were gathered by the Pediatric Department of the University Hospital of Alexandroupolis, Greece during the period from 2008 to 2010. A group of 148 patients were diagnosed for asthma and were studied prospectively from the 7th to the 14th year of age. The history of each case was obtained by questionnaire and 36 patients were removed from the study due to missing values. A subsequent group of 33 children was used for validation and predictability examination of the ridge regression models. This group of preschool children is used to predict asthma persistence in school age through logistic ridge regression models with different ridge parameters. The new dataset has 18 prognostic factors which have been derived by previous studies [34-36] and they are described in Table I. The 18 variables inevitably will become 23, as the factor "seasonal symptoms" has to become a dummy variable. The encoding of the prognostic factor "seasonal symptoms" is presented in Table II.

TABLE I

Category	Prognostic Factors
Demographic	Age, height, weight, waist's perimeter
Bronchiolitis episodes	Until 3rd year, between 3rd – 5th year
Symptoms	Wheezing, cough, allergic rhinitis, allergic conjunctivitis, dyspnea, congestion, runny nose, seasonal symptoms
Pharmaceutical therapy	Antileukotriene, antihistamine, corticosteroids inhaled
Asthma	Diagnosis of asthma (dependent variable), Treatment

The 18 used prognostic factors.

TABLE II

1 (none)	2 (Winter)	3 (Autumn)	4 (Spring)	5 (Summer)	6 (>2seasons)

The encoding of "seasonal symptoms".

B. Logistic Ridge Regression

In this section the implementation of the logistic ridge regression is presented.

The logistic regression model with the use of the logit link function is:

$$log\left(\frac{p_i}{1-p_i}\right) = b\boldsymbol{x}_i, \quad (1)$$

which is equivalent to:

$$p_i = \frac{\exp(b\boldsymbol{x}_i)}{\{1+\exp(b\boldsymbol{x}_i)\}}, \quad (2)$$

where b is the parameter vector and $\boldsymbol{x}_i$ is a data matrix of explanatory variables.

In order to estimate b the maximum likelihood method is applied. The estimates of the parameters b_j, j=1,…,k are obtained by maximizing the log – likelihood which is:

$$l(b|y) = \log L(b|y), \quad (3)$$

$$l(b|y) = \sum_{i=1}^{n}\left[y_i \log\left[\frac{1}{1+\exp(-x_i^T b)}\right] + (1-y_i)\log\left[1-\frac{1}{1+\exp(-x_i^T b)}\right]\right] \quad (4)$$

As it was mentioned before when multicollinearity exists, in order to obtain more stable parameter estimates the logistic ridge regression is used. In order to improve further the estimation procedure, a penalized likelihood function is implemented given by [17]:

$$l^{\lambda}(b|y) = l(b|y) - \lambda\|b\|^2 = l(b|y) - \lambda R, \quad (5)$$

where, R is a penalty term of the following form [33]:

$$R = \sum_{j=0}^{k-1}(b_{j+1}-b_j)^2. \quad (6)$$

It is obvious that in this approach a penalty term is included that contains the ridge parameter λ. The penalty term improves the properties of the estimated parameters b in (5). The computation of the estimates of the penalized parameters $\hat{b}^{\lambda}$ is based on the Newton – Raphson's iterative algorithm. In order to be able to use the relation (5), a transformation of the parameters of the unrestricted logistic model is required.

Therefore:

$$b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} = \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_k z_{ik}, \quad (7)$$

where

$$\gamma_1 = b_1, \dots,\ \gamma_j = b_j - b_{j-1}, \quad j = 2, \dots, k \quad (7.1)$$

and $z_{ij} = \sum_{u=j}^{k} x_{iu}$ (7.2). Thus, the penalized likelihood function can be written as follows:

$$l^{\lambda}(\gamma|y) = l(\gamma|y) - \lambda\|\gamma\|^2\,. \quad (8)$$

Applying the procedure described in [17,18] we obtain

$$\hat{\gamma}^{\lambda} = \{\Omega(\gamma) + 2\lambda I\}^{-1}\{U(\gamma) + \gamma\Omega(\gamma)\}. \quad (9)$$

where $\Omega(\gamma) = \boldsymbol{z^T W z},\ U(\gamma) = \sum_{i=1}^{n} z_i\{y_i - p_i\}$ and γ are the estimated coefficients of the unrestricted model.

C. The ridge parameter

When a ridge regression model is implemented, the choice of the ridge parameter is of great importance. The most classical and usually used ridge parameter is the one proposed by Hoerl and Kennard [2-3],

$$\lambda_1 = \lambda_{HK1} = \frac{s^2}{\hat{a}_{max}{}^2}$$

where $\hat{a}_{max}{}^2$ is the maximum element of δb_{ML}, where δ is the eigenvector of $\boldsymbol{X'WX}$, b_{ML} are the estimates of the unrestricted maximum likelihood and

$$s^2 = \frac{(y-\hat{y})^T(y-\hat{y})}{n-p-1}$$

Another version of the previous ridge parameter is proposed by [14]:

$$\lambda_2 = \lambda_{SRW} = \frac{1}{\hat{a}_{max}{}^2},$$

Moreover, two other ridge parameters discussed in [8] are given by:

$$\lambda_3 = \lambda_{GM} = \frac{s^2}{\left(\prod_{i=1}^{p}\hat{a}_i{}^2\right)^{\frac{1}{p}}},$$

and

$$\lambda_4 = \lambda_{MED} = Median\left(\frac{s^2}{\hat{a}_i^{\,2}}\right), i = 1,2,\dots,p$$

Also, Alkhamisi et.al proposed the following ridge parameter [12]:

$$\lambda_5 = \lambda_{AL} = \max\left[\frac{t_i s^2}{(n-p-1)s^2 + t_i \hat{a}_i^{\,2}}\right],$$

where t_i are the eigenvalues of $\boldsymbol{X'X}$ matrix.

Furthermore, one way of selecting an appropriate Ridge Parameter is the process of Cross Validation. In this direction it is possible to perform an estimate of the mean squared error of the cross validation set, which can be minimized to obtain the Ridge parameter. The prediction errors that are widely used in accordance with [17], are:

(a) The mean classification error

$$\mathrm{MCE_{cv}} = \frac{1}{n}\sum_i \mathrm{Y_i}\left[\hat{\mathrm{p}}_\mathrm{i}(\mathrm{X_i}) < \frac{1}{2}\right] + (1-\mathrm{Y_i})\left[\hat{\mathrm{p}}_{(\mathrm{i})}(\mathrm{X_i}) > \frac{1}{2}\right] + \frac{1}{2}\left[\hat{\mathrm{p}}_{(\mathrm{i})}(\mathrm{X_i}) = \frac{1}{2}\right]$$

where if the proposition inside [ ] is true then it takes the value 1 and if it is false takes the value 0.

(b) The mean squared error

$$MSE_{cv} = \frac{1}{n}(\sum_i \left\{Y_i - \hat{p}_{(i)}(X_i)\right\}^2$$

(c) and the mean minus log – likelihood

$$MML_{cv} = -\frac{1}{n}\sum_i \left[Y_i \log \hat{p}_{(i)}(X_i) + (1-Y_i)\log\left\{1-\hat{p}_{(i)}(X_i)\right\}\right]$$

D. Residuals and bootstrapping

A usual problem that occurs in logistic regression is the validity examination of the model based on the residuals. In the case of logistic regression with binary response, the Pearson residuals which are defined by $r_{p,i} = (y_i - \hat{p}_i)/\sqrt{\hat{p}_i(1-\hat{p}_i)}$, $i = 1,\dots,n$ and the deviance residuals which are defined by, $r_D = sign\{y_i - \hat{p}_i\}$ are far from normal and as a result are not capable to give us any information about the validity of the model. More details about the residuals and their use are presented in [30].

For that reason the randomized quantile residuals proposed by Dunn and Smith are used [19]. The randomized quantile residuals are defined as follows:

Let $F(y_i; p_i) = P(Y_i \le y_i) = \sum_{m=0}^{\lfloor y_i \rfloor} p_i^m (1-p_i)^{1-m}$ be the cumulative binomial distribution of the ith binary response, and $\lfloor y_i \rfloor$ is the greatest integer less than or equal to y_i, i.e. the 'floor' under y_i. Let also

$$a_i = \lim_{y \uparrow y_i} F(y; \hat{p_i}) \text{ and } b_i = F(y; \hat{p_i}).$$

Then the randomized quantile residuals for a logistic regression model are defined by

$$r_{rq,i} = \Phi^{-1}\{u\}, \qquad (11)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, and u_i is a uniform random variable on the interval $(a_i, b_i]$.

These residuals [19] can be used for any discrete distributed response. Thus, the validity of the model can now be tested by using goodness of fit tests for the normality of $r_{rq,i}$. A very commonly used method to test the null hypothesis that the randomized quantile residuals follow a standard normal distribution i.e. $\boldsymbol{r_{rq}} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I})$ is the Anderson – Darling test [25].

Also the Q-Q plot of the randomized quantile can be a mean for checking the validity of the model. A method for constructing pointwise a $\times$ 100% rejection regions around the Q-Q plot of any random sample is proposed in [18] by using bootstrapping.

III. Results

The correlations between some variables are very strong and statistically significant, indicating the presence of multicollinearity. The condition indices also reveal that multicollinearity exists [18,29].

Thus the logistic ridge regression is applied for a ridge parameter λ=0 to λ=1. Furthermore it is important to mention that when collinearity exists there is always a model for λ>0 for which the MSE is less than the MSE of the unrestricted model [3,17].

For the calculation of p – values the following statistic is used:

$$T_\lambda = \frac{\hat{b}_j^\lambda}{se\left(\hat{b}_j^\lambda\right)} \qquad (10)$$

The standard errors are obtained by a bootstrap procedure using the randomized quantile residuals that is described in [18]. Thereafter we assume that under the null hypothesis $T_\lambda \sim N(0,1)$ to test the significance of the estimated ridge coefficients [20].

The results of using different approaches for the ridge parameter calculation are shown in TABLE III-X (Appendix).

Now, we would like to examine the performance of these models in new real data. These new data refer to 33 new patients and were collected also by questionnaire in a period after 2010.

Based on the equation:

$$\hat{p}_{ridge} = \frac{1}{1 + \exp(-X_{new} * \hat{b}_{ridge})},$$

a prediction for the diagnosis of a new patient can be found. The positive predicted value, the negative predicted value and the accuracy of this model are estimated using false positive (FP), false negative (FN), true positive (TP), and true negative (TN) values. The positive predictive value (PPV) of a test is defined as the proportion of people with a positive test result who actually have the disease. The negative predictive value (NPV) of a test is the proportion of people with a negative test result who do not have disease [21]. The test set consists of the new 33 patients and the 11 patients which were used for the cross – validation test.

$$Positive\ Pred.Value = \frac{N_{TP}}{N_{TP}+N_{FP}} \times 100,$$
$$Negative\ Pred.Value = \frac{N_{TN}}{N_{TN}+N_{FN}} \times 100, \qquad (12)$$
$$Accuracy = \frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FP}+N_{FN}} \times 100.$$

All the above are statistical measures of the performance of a binary classification test.

Those measures are very useful and give us important information about a patient. For example if a PPV of a disease prediction model is 90% then a patient with a positive test has a chance of 90% having the particular disease [21].

In the following TABLE XI are described all the statistical measures for the performance of the models including the mean squared error of the data that were used for fitting the models.

TABLE XI

Ridge Parameters	MSE	Accuracy (%)	PPV (%)	NPV (%)	AIC
λ_1=0.000456	0.0813	72.73	81.82	63.64	574.5633
λ_2=0.0159	0.0857	86.36	88.46	83.33	319.4652
λ_3=0.0018	0.0849	75	82.60	66.67	487.6483
λ_4=0.012	0.0847	77.27	83.33	70	512.1119
λ_5=0.0541	0.0846	84.09	88	78.95	409.9047
λ_6=0.0261	0.0884	93.18	96.15	88.89	276.1508
λ_7=0.0160	0.0850	86.36	88.46	83.33	341.1905
λ_8=0.0123	0.0858	86.36	88.46	83.33	319.0086

The results show that the most significant explanatory variable that appears in all models is the waist's perimeter. Waist's perimeter is studied and also presented as a significant variable in [22-23] and that enhances the fact that there is a strong relation between asthma and obesity.

One other interesting matter is that the model of TABLE III (Appendix) has many different significant variables compared to the other models but it also has the smallest predictive accuracy. That is explained if we perform a validity test using the QQ-plot of the randomized quantile residuals with the use of the method described in [18]. This method uses the bootstrap resampling of randomized quantile residuals so we can calculate the 5% rejection regions around the QQ-plot. This is implemented because the large number of the estimated parameters adds extra uncertainty.

More specific, we obtain estimates of $\hat{p}_i$, and randomized quantile residuals with the use of logistic ridge regression. Then we apply the bootstrap 2000 times in randomized quantile residuals and then we apply again the logistic ridge regression with an assumption made in [24,37] 2000 times using as response the summations $\widehat{\boldsymbol{p}}^{\boldsymbol{T}} + \boldsymbol{r}^{\boldsymbol{T}}_{\boldsymbol{rq,t}}$. Finally from the 2000 sets of estimated response variables $\hat{p}_t, t = 1, \dots, 2000$, we calculate 2000 new sets of randomized quantile residuals which allows us to construct $a \times 100\%$ rejection regions around the Q-Q plot of the randomized quantile residuals.

Here it is important to mention that the standard errors of the estimated coefficients $\hat{b}^{\lambda}$ were obtained by finding the standard deviation of the 2000 bootstrapped samples $\hat{b}^{\lambda}_1, \dots, \hat{b}^{\lambda}_{23}$.

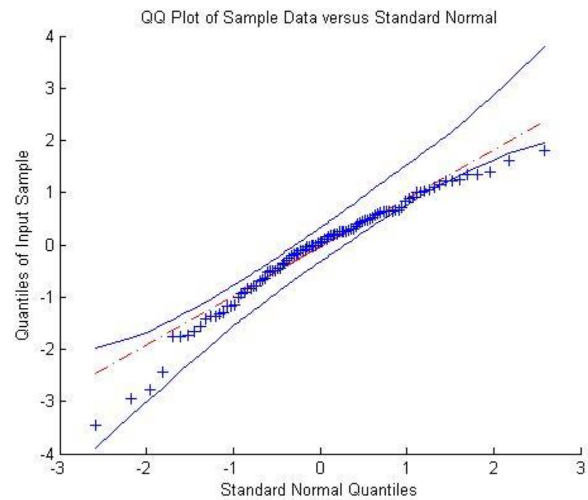

Fig. 1 QQ Plot of randomized quantile residuals of LRR model with the use of $\lambda_1 = \lambda_{HK1}$ versus Standard Normal

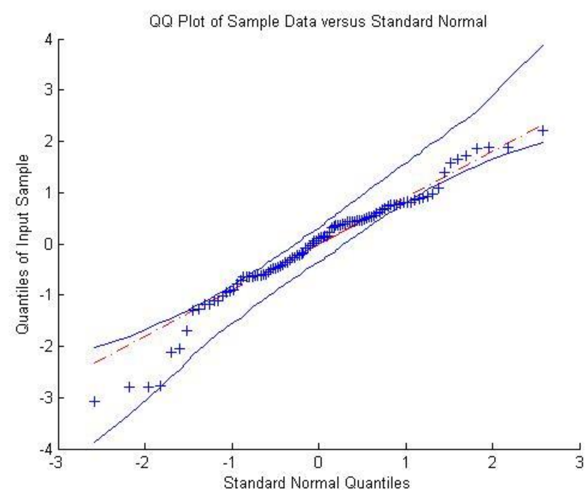

Fig. 2 QQ Plot of randomized quantile residuals of LRR model with the use of $\lambda_3 = \lambda_{GM}$ versus Standard Normal

Figures 1 and 2 show the Q-Q plot of the randomized quantile residuals of the fitted logistic ridge model with λ_{HK1} and λ_{GM} denoted both with +. The 5% rejection regions were computed by the procedure described above after 2000 bootstrap simulations. It is observed that 7 (6.93%) and 9 (8.91%) of the 101 residuals lie outside the 5% rejection regions respectively and generally the Q-Q plots present some deviations from normality.

In addition, the Anderson-Darling test gives the value 1.0559 with a p-value 0.0084 and 1.0738 with a p-value 0.0076 respectively. Therefore, the null hypothesis that the randomized quantile residuals follow an approximate standard normal distribution must be rejected. This suggests that the fitted models are invalid. The rest of the models are valid and this is proved with the use of the same procedure for checking the validity.

It is also interesting the fact that the models which have smaller MSE have larger accuracy and the opposite. This is probably a result of overfitting. This is also the reason that the models with smaller MSEs are not valid according to our validity test with the use of randomized quantile residuals.

Finally we may use the Akaike Information Criterion in order to determine which model is the best for asthma persistence prediction. AIC is defined as [28]:

$$AIC = nlog(RSS) + 2df$$

where RSS is the residual sum of squares and df is defined as:

$$df = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

where d_j, $j = 1, \dots, p$ is the vector of singular values after the singular value decomposition (SVD) of $\boldsymbol{X}$ [28]

The Akaike Information criterion of each model is presented in TABLE XI.

IV. Conclusion

A complete ridge regression study demands the use of many different ridge parameters and a number of criteria in order to obtain the best model. In this paper the proposed models for ridge parameters $\lambda_2, \lambda_5, \lambda_6, \lambda_7, \lambda_8$ exhibit high accuracy and small mean squared errors. The best model according to four different criteria is the one in which the ridge parameter is calculated by minimizing the mean squared error through cross – validation. The values of the four criteria are:

$Accuracy$: 93.18%
PPV: 96.15%,
NPV: 88.89% and
AIC: 276.15.

For future research, a very interesting study would be to make a comparison of the ridge regression models for asthma prediction, with other methods used to deal with multicollinearity such as partial least squares regression, principal component analysis and Bayesian logistic regression. A comparison with artificial intelligence methods such as artificial neural networks [27] can also be included.

V.APPENDIX

TABLE III

Covariates	Estimates			
	Parameter Estimates	***Standard Errors***	T_λ	***p-values***
Age	-0.1216	0.1825	-0.6664	0.5052
Treatment	-0.5972	0.9194	-0.6496	0.5160
Corticosteroids inhaled	1.2599	0.9495	1.3269	0.1845
Antileukotriene	-0.5388	1.1484	-0.4692	0.6390
Antihistamine	-1.9814	1.2809	-1.5469	0.1219
Height	-3.1837	2.1750	-1.4638	0.1433
Weight	0.1107	0.0394	2.8084	0.0050
Waist's perimeter	-0.1331	0.0348	-3.8224	0.0001
Allergic rhinitis	0.7402	1.0964	0.6751	0.4996
Allergic conjunctivitis	-2.8051	1.3000	-2.1579	0.0309
Runny nose	2.4870	1.1636	2.1373	0.0326
Congestion	0.9241	1.0703	0.8633	0.3879
Cough	1.4526	1.1453	1.2683	0.2047
Wheezing	2.4733	1.1754	2.1042	0.0354
Dyspnea	1.1429	1.0215	1.1189	0.2632
Seasonal symptoms (none)	5.7369	2.1801	2.6315	0.0085
Seasonal symptoms (winter)	7.7949	2.3360	3.3369	0.0008
Seasonal symptoms (autumn)	5.2621	2.3813	2.2098	0.0271
Seasonal symptoms (spring)	8.5665	2.4856	3.4464	0.0006
Seasonal symptoms (summer)	4.2801	2.5134	1.7029	0.0886
Seasonal symptoms (>2 seasons)	6.9742	2.3594	2.9559	0.0031
Bronchiolitis episodes until 3rd year	-0.1539	0.1561	-0.9858	0.3242
Bronchiolitis episodes b/w 3rd – 5th year	0.2086	0.1350	1.5447	0.1224

TABLE III: The logistic ridge regression model for λ_1 which is equal to 0.000456.

TABLE IV

Covariates	Estimates			
	Parameter Estimates	***Standard Errors***	T_λ	***p-values***
Age	0.0250	0.1502	0.1665	0.8678
Treatment	0.4897	0.6046	0.8099	0.4180
Corticosteroids inhaled	0.9499	0.6217	1.5279	0.1265
Antileukotriene	-0.4523	0.8828	-0.5123	0.6084
Antihistamine	-0.0078	0.9693	-0.0081	0.9936
Height	0.7008	1.2574	0.5573	0.5773
Weight	-0.0010	0.0319	-0.0301	0.9760
Waist's perimeter	-0.0759	0.0288	-2.6364	0.0084
Allergic rhinitis	-0.0272	0.8375	-0.0325	0.9741
Allergic conjunctivitis	-0.9199	0.9208	-0.9991	0.3177
Runny nose	0.6202	0.8670	0.7154	0.4744

Covariates	Estimates			
	Parameter Estimates	***Standard Errors***	T_λ	***p-values***
Congestion	1.1530	0.8020	1.4378	0.1505
Cough	1.7631	0.8653	2.0375	0.0416
Wheezing	1.7896	0.8625	2.0750	0.0380
Dyspnea	1.1127	0.8009	1.3893	0.1647
Seasonal symptoms (none)	1.4128	1.2993	1.0874	0.2769
Seasonal symptoms (winter)	1.4186	1.4910	0.9514	0.3414
Seasonal symptoms (autumn)	0.3989	1.4973	0.2664	0.7899
Seasonal symptoms (spring)	1.2300	1.6099	0.7640	0.4449
Seasonal symptoms (summer)	0.2360	1.6090	0.1467	0.8834
Seasonal symptoms (>2 seasons)	1.4584	1.4819	0.9842	0.3250
Bronchiolitis episodes until 3rd year	-0.2146	0.1337	-1.6047	0.1086
Bronchiolitis episodes b/w 3rd – 5th year	0.1461	0.1137	1.2852	0.1987

TABLE IV: The logistic ridge regression model for λ_2 which is equal to 0.0159.

TABLE V

Covariates	Estimates			
	Parameter Estimates	***Standard Errors***	T_λ	***p-values***
Age	-0.1124	0.2116	-0.5315	0.5951
Treatment	-0.1510	1.3318	-0.1134	0.9097
Corticosteroids inhaled	1.2057	1.4395	0.8376	0.4023
Antileukotriene	-0.6129	1.3353	-0.4590	0.6463
Antihistamine	-1.4135	1.5358	-0.9204	0.3574
Height	-0.1232	3.1444	-0.0392	0.9687
Weight	0.0628	0.0437	1.4396	0.1500
Waist’s perimeter	-0.1193	0.0372	-3.2080	0.0013
Allergic rhinitis	0.4453	1.3051	0.3412	0.7329
Allergic conjunctivitis	-2.0926	1.5382	-1.3604	0.1737
Runny nose	1.8047	1.3142	1.3732	0.1697
Congestion	1.0902	1.1989	0.9093	0.3632
Cough	1.7637	1.3620	1.2950	0.1953
Wheezing	2.2655	1.3315	1.7015	0.0888
Dyspnea	0.9703	1.1591	0.8371	0.4025
Seasonal symptoms (none)	2.8558	2.8203	1.0126	0.3113
Seasonal symptoms (winter)	4.0640	2.8922	1.4051	0.1600
Seasonal symptoms (autumn)	2.0847	3.2066	0.6501	0.5156
Seasonal symptoms (spring)	4.6342	3.0200	1.5345	0.1249
Seasonal symptoms (summer)	1.2966	3.0580	0.4240	0.6716
Seasonal symptoms (>2 seasons)	3.6561	2.9673	1.2321	0.2179
Bronchiolitis episodes until 3rd year	-0.1791	0.1597	-1.1214	0.2621
Bronchiolitis episodes b/w 3rd – 5th year	0.1968	0.1389	1.4172	0.1564

TABLE V: The logistic ridge regression model for λ_3 which is equal to 0.0018.

TABLE VI

Covariates	Estimates			
	Parameter Estimates	***Standard Errors***	T_λ	***p-values***
Age	-0.1227	0.2039	-0.6015	0.5475
Treatment	-0.3004	1.4298	-0.2101	0.8336
Corticosteroids inhaled	1.2330	1.5874	0.7768	0.4373
Antileukotriene	-0.5896	1.3675	-0.4312	0.6663
Antihistamine	-1.6547	1.7057	-0.9701	0.3320
Height	-0.6974	3.1056	-0.2245	0.8223
Weight	0.0760	0.0465	1.6328	0.1025
Waist’s perimeter	-0.1243	0.0394	-3.1576	0.0016
Allergic rhinitis	0.5572	1.3072	0.4262	0.6699
Allergic conjunctivitis	-2.3310	1.5145	-1.5391	0.1238
Runny nose	2.0192	1.3832	1.4598	0.1443
Congestion	1.0387	1.2286	0.8454	0.3979
Cough	1.6524	1.2987	1.2723	0.2033
Wheezing	2.3673	1.3991	1.6921	0.0906
Dyspnea	0.9937	1.2202	0.8144	0.4154
Seasonal symptoms (none)	3.4249	2.9862	1.1469	0.2514
Seasonal symptoms (winter)	4.9009	3.1197	1.5709	0.1162
Seasonal symptoms (autumn)	2.7220	3.4452	0.7901	0.4295
Seasonal symptoms (spring)	5.5520	3.1196	1.7797	0.0751
Seasonal symptoms (summer)	1.8582	3.2042	0.5799	0.5620
Seasonal symptoms (>2 seasons)	4.3439	3.0055	1.4453	0.1484
Bronchiolitis episodes until 3rd year	-0.1722	0.1712	-1.0059	0.3145
Bronchiolitis episodes b/w 3rd – 5th year	0.2018	0.1475	1.3678	0.1714

TABLE VI: The logistic ridge regression model for λ_4 which is equal to 0.012.

TABLE VII

Covariates	Estimates			
	Parameter Estimates	*Standard Errors*	T_λ	*p-values*
Age	-0.0540	0.2041	-0.2645	0.7914
Treatment	0.2506	1.2937	0.1937	0.8464
Corticosteroids inhaled	1.0861	1.4304	0.7593	0.4477
Antileukotriene	-0.6135	1.3484	-0.4550	0.6491
Antihistamine	-0.5924	1.4285	-0.4147	0.6784
Height	0.6537	3.1017	0.2108	0.8331
Weight	0.0278	0.0470	0.5926	0.5534
Waist's perimeter	-0.0999	0.0394	-2.5339	0.0113
Allergic rhinitis	0.1472	1.2997	0.1132	0.9098
Allergic conjunctivitis	-1.4431	1.5195	-0.9497	0.3423
Runny nose	1.1841	1.3081	0.9052	0.3654
Congestion	1.1818	1.1882	0.9946	0.3199
Cough	1.9106	1.4340	1.3324	0.1827
Wheezing	1.9913	1.3684	1.4552	0.1456
Dyspnea	0.9835	1.1951	0.8230	0.4105
Seasonal symptoms (none)	1.9139	2.9094	0.6578	0.5106
Seasonal symptoms (winter)	2.4033	2.9500	0.8147	0.4153
Seasonal symptoms (autumn)	0.9678	3.3513	0.2888	0.7727
Seasonal symptoms (spring)	2.5807	3.1280	0.8250	0.4094
Seasonal symptoms (summer)	0.4833	3.1612	0.1529	0.8785
Seasonal symptoms (>2 seasons)	2.3315	3.0638	0.7610	0.4467
Bronchiolitis episodes until 3rd year	-0.2015	0.1689	-1.1927	0.2330
Bronchiolitis episodes b/w 3rd – 5th year	0.1767	0.1427	1.2381	0.2157

TABLE VII: The logistic ridge regression model for λ_5 which is equal to 0.0541.

TABLE VIII

Covariates	Estimates			
	Parameter Estimates	*Standard Errors*	T_λ	*p-values*
Age	0,059821	0,1221	0,4900	0,6241
Treatment	0,531238	0,4817	1,1028	0,2701
Corticosteroids inhaled	0,889768	0,4942	1,8002	0,0718
Antileukotriene	-0,32763	0,5650	-0,5799	0,5620
Antihistamine	0,111097	0,6674	0,1665	0,8678
Height	0,600211	0,7353	0,8163	0,4143
Weight	-0,01082	0,0276	-0,3925	0,6947
Waist's perimeter	-0,06579	0,0216	-3,0455	0,0023
Allergic rhinitis	-0,06907	0,5733	-0,1205	0,9041
Allergic conjunctivitis	-0,72709	0,5819	-1,2494	0,2115
Runny nose	0,429069	0,6068	0,7071	0,4795
Congestion	1,096703	0,5424	2,0220	0,0432
Cough	1,640577	0,5871	2,7942	0,0052
Wheezing	1,719255	0,5874	2,9271	0,0034
Dyspnea	1,18429	0,5962	1,9865	0,0470
Seasonal symptoms (none)	1,26928	0,7360	1,7246	0,0846
Seasonal symptoms (winter)	1,148824	0,8505	1,3507	0,1768
Seasonal symptoms (autumn)	0,273617	0,8385	0,3263	0,7442
Seasonal symptoms (spring)	0,856916	0,9088	0,9429	0,3457
Seasonal symptoms (summer)	0,216933	0,8830	0,2457	0,8059
Seasonal symptoms (>2 seasons)	1,15655	0,8174	1,4149	0,1571
Bronchiolitis episodes until 3rd year	-0,21639	0,1089	-1,9866	0,0470
Bronchiolitis episodes b/w 3rd – 5th year	0,132402	0,0932	1,4212	0,1553

TABLE VIII: The parameter estimates of the logistic ridge model for the minimum MSEcv. The minimum MSEcv is derived for λ_6=0.0261 and is equal to 0.034.

TABLE IX

Covariates	Estimates			
	Parameter Estimates	*Standard Errors*	T_λ	*p-values*
Age	0.0062	0.1357	0.0460	0.9633
Treatment	0.4527	0.7109	0.6368	0.5243
Corticosteroids inhaled	0.9812	0.7369	1.3315	0.1830
Antileukotriene	-0.5055	0.8132	-0.6216	0.5342
Antihistamine	-0.1040	0.8708	-0.1195	0.9049
Height	0.7314	1.3985	0.5230	0.6010
Weight	0.0050	0.0302	0.1646	0.8693
Waist's perimeter	-0.0815	0.0262	-3.1141	0.0018
Allergic rhinitis	0.0021	0.7788	0.0027	0.9978
Allergic conjunctivitis	-1.0293	0.8743	-1.1774	0.2391
Runny nose	0.7381	0.8086	0.9128	0.3613
Congestion	1.1721	0.7242	1.6184	0.1056
Cough	1.8179	0.8107	2.2423	0.0249
Wheezing	1.8300	0.8252	2.2176	0.0266
Dyspnea	1.0757	0.7490	1.4362	0.1509
Seasonal symptoms (none)	1.5054	1.3621	1.1052	0.2691
Seasonal symptoms (winter)	1.5978	1.4433	1.1070	0.2683
Seasonal symptoms (autumn)	0.4939	1.5525	0.3181	0.7504
Seasonal symptoms (spring)	1.4782	1.5721	0.9403	0.3471
Seasonal symptoms (summer)	0.2629	1.5506	0.1696	0.8654

Covariates	Estimates			
	Parameter Estimates	*Standard Errors*	T_λ	*p-values*
Seasonal symptoms (>2 seasons)	1.6343	1.4832	1.1018	0.2705
Bronchiolitis episodes until 3rd year	-0.2127	0.1269	-1.6769	0.0936
Bronchiolitis episodes b/w 3rd – 5th year	0.1537	0.1102	1.3940	0.1633

TABLE IX: The parameter estimates of the logistic ridge model for the minimum MMLcv. The minimum MMLcv is derived for λ_7=0.0160 and is equal to 0.1693.

TABLE X

Covariates	Estimates			
	Parameter Estimates	*Standard Errors*	T_λ	*p-values*
Age	-0.1216	0.1771	0.1440	0.8855
Treatment	-0.5972	0.9529	0.5148	0.6067
Corticosteroids inhaled	1.2599	1.0058	0.9436	0.3454
Antileukotriene	-0.5388	1.0843	-0.4157	0.6777
Antihistamine	-1.9814	1.1873	-0.0047	0.9963
Height	-3.1837	2.3980	0.2918	0.7704
Weight	0.1107	0.0412	-0.0271	0.9784
Waist's perimeter	-0.1331	0.0333	-2.2774	0.0228
Allergic rhinitis	0.7402	1.0491	-0.0266	0.9788
Allergic conjunctivitis	-2.8051	1.1948	-0.7676	0.4427
Runny nose	2.4870	1.0590	0.5828	0.5600
Congestion	0.9241	0.9924	1.1612	0.2456
Cough	1.4526	1.0918	1.6134	0.1067
Wheezing	2.4733	1.0963	1.6315	0.1028
Dyspnea	1.1429	0.9810	1.1354	0.2562
Seasonal symptoms (none)	5.7369	2.3044	0.6121	0.5405
Seasonal symptoms (winter)	7.7949	2.3937	0.5908	0.5547
Seasonal symptoms (autumn)	5.2621	2.6060	0.1522	0.8790
Seasonal symptoms (spring)	8.5665	2.5260	0.4845	0.6280
Seasonal symptoms (summer)	4.2801	2.5459	0.0925	0.9263
Seasonal symptoms (>2 seasons)	6.9742	2.4144	0.6022	0.5471
Bronchiolitis episodes until 3rd year	-0.1539	0.1495	-1.4359	0.1510
Bronchiolitis episodes b/w 3rd – 5th year	0.2086	0.1260	1.1582	0.2468

TABLE X: The parameter estimates of the logistic ridge model for the minimum MCEcv. The minimum MCEcv is derived for λ_8= 0.0123 and is equal to 0.

REFERENCES

[1] R. Frisch, *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: University Institute of Economics, Publication no. 5, 1934.

[2] A.E. Hoerl and R. W. Kennard, "Ridge Regression: Applications to nonorthogonal problems," *Technometrics 1*, vol. 12, no. 6, 1970.

[3] A.E. Hoerl and R.W. Kennard, "Ridge Regression: Biased estimates for nonorthogonal problems," *Technometrics*, vol. 12, no.55, pp.55–67 1970.

[4] McDonald, G. C., Galarneau, D. I. A Monte Carlo evaluation of some ridge-type estimators. J. *Amer. Statist. Assoc*. vol.70, pp.407–416, 1975

[5] Lawless, J. F., Wang, P.. A simulation study of ridge and other regression estimators. Commun. Statist. Theor. Meth. vol.5, pp.307–323, 1976.

[6] Saleh, A. K. M., Kibria, B. M. G. Performance of some new preliminary test ridge regression and their properties. *Commun. Statist. Theor. Meth*. vol.22, pp.2747–2764, 1993.

[7] Haq, M. S., Kibria, B. M. G. A shrinkage estimator for the restricted linear regression model: Ridge regression approach. J. Appl. Statist. Sci. 3:301–316, 1996.

[8] Kibria, B. M. G. Performance of some new ridge regression estimators. Commun. Statist. Theor. Meth. 32:419–435, 2003.

[9] Kibria, B. M. G. Performance of the shrinkage preliminary test ridge regression estimators based on the conflicting of W, LR and LM tests. J. Statist. Computat. Simul. 74:793–810, 2004.

[10] Kibria, B. M. G., Saleh, A. K. M. E. Performance of positive rule estimator in the ill-conditioned gaussian regression model. Calcutta Statist. Assoc. Bull. 55:211–241, 2004.

[11] Khalaf, G., Shukur, G. Choosing ridge parameters for regression problems. Commun. Statist. Theor. Meth. 34:1177–1182, 2005.

[12] Alkhamisi, M. A., Khalaf, G., Shukur, G. Some modifications for choosing ridge parameter. Commun. Statist. Theor. and Meth. 35:1–16, 2006.

[13] Muniz, G., Kibria, B. M. G. On some ridge regression estimators: An empirical comparison. Commun. Statist. Simul. Computat. 38:621–630, 2009.

[14] R. L. Schaefer, L.D. Roi, and R.A. Wolfe, "A ridge logistic estimator, " *Communications in Statistics - Theory and Methods*, vol. 13, no. 1, pp. 99–113, 1984.

[15] Hoerl, A. E., Kennard, R. W., Baldwin, K. F. Ridge regression: Some simulation. Commun. Statist. Theor. Meth. 4:105–123, 1975.

[16] Kristofer Månsson & Ghazi Shukur, On Ridge Parameters in Logistic Regression, *Communications in Statistics - Theory and Methods*, 40:18, 3366-3381, 2011, DOI: 10.1080/03610926.2010.500111.

[17] S. Le Cessie and J. C. Van Houwlingen, "Ridge Estimators in Logistic Regression," *Journal of the Royal Statistical Society*, Series C (Applied Statistics), vol. 41, no. 1, pp. 191–201, 1992.

[18] Ioannis I. Spyroglou, Eleni A. Chatzimichail, E.N. Spanou, E. Paraskakis, and Alexandros G. Rigas, "Ridge regression and bootstrapping in asthma prediction," in New Developments in Pure and Applied Mathematics INASE Conference proceedings (MMSSE '15), Vienna, Austria, pp. 44-48, March 2015.

[19] P. Dunn and G. K. Smyth, "Randomized Quantile Residuals, " *J. Computat. Graph. Statist*, vol. 5, pp. 236–244, 1996.

[20] Cule et al., "Significance testing in ridge regressionfor genetic data," *BMC Bioinformatics*, 12:372, 2011.

[21] Akobeng, A. K.. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta paediatrica, 96(3), 338-341, 2007.

[22] McGarry ME, Castellanos E, Thakur N, Oh SS, Eng C, Davis A, Meade K, LeNoirMA, Avila PC, Farber HJ, Serebrisky D, Brigino-Buenaventura E, Rodriguez-Cintron W, Kumar R, Bibbins-Domingo K, Thyne SM, Sen S, Rodriguez-Santana JR, Borrell LN,Burchard EG. Obesity and Bronchodilator Response in African-American and Hispanic Children and Adolescents with Asthma. Chest, 2015, doi:10.1378/chest.14-2689.

[23] Spathopoulos D, Paraskakis E, Trypsianis G, Tsalkidis A, Arvanitidou V,Emporiadou M, Bouros D, Chatzimichael A. The effect of obesity on pulmonary lung function of school aged children in Greece. Pediatr Pulmonol., 44(3):273-80, 2009.

[24] H. Friedl and N. Tilg , "Variance estimates in logistic regression using the bootstrap, " *Communications in Statistics - Theory and Methods*, vol. 24, no. 2, pp. 473–486, 1995.

[25] T.W. Anderson and D.A. Darling, “Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes,” *The Annals of Mathematical Statistics*, vol. 23, no. 2, pp. 193–212, 1952.
[26] B. Efron and R.J. Tibshirani, “An Introduction to the Bootstrap, ” (Chapman & Hall, New York), 1993.
[27] E. Chatzimichail, E. Paraskakis, and A. Rigas, “Predicting Asthma Outcome Using Partial Least Square Regression and Artificial Neural Networks,” *Advances in Artificial Intelligence*, vol. 2013, Article ID 435321, 7 pages, 2013. doi:10.1155/2013/435321
[28] Hastie Trevor, Tibshirani Robert, and Friedman Jerome, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Stanford, California: Springer Series in Statistics, 2008.
[29] E. Lesaffre, E. and B.D. Marx, “Collinearity in Generalized Linear Regression,”*Communications in Statistics – Theory and Methods*, vol. 22, no. 7, pp. 1933–1952, 1993.
[30] D.A. Pierce and D.W. Schafer, “Residuals in Generalized Linear Models,” *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 977–986, 1986.
[31] Collett D., *Modelling Binary Data*, Second Edition. Chapman and Hall, Boca Raton, 2003
[32] D.A. Freedman , “ Bootstrapping regression models, ” Ann. Statist., vol. 9, pp. 1218–1228, 1981.
[33] D.R. Brillinger, K.A. Lindsay, and J.R. Rosenberg, “Combining frequency and time domain approaches to systems with multiple spike train input and output,” *Biological Cybernetics*, vol. 100, pp. 459–474, 2009.
[34] C. Porsbjerg, M.L. von Linstow, C. Ulrik, S. Nepper-Christensen, V. Backer, “Risk factors for onset of asthma: a 12-year prospective follow-up study,” Chest, vol. 129, no. 2, pp. 309–16, 2006.
[35] N. N. Hansel, E. C. Matsui, R. Rusher, M. C. McCormack, J. Curtin-Brosnan, R. D. Peng, D. Mazique, P. N. Breysse, G. B. Diette, “Predicting future asthma morbidity in preschool inner-city children, ” Journal of Asthma, vol. 48, no.8, pp. 797–803, 2011.
[36] A. Bush, “Diagnosis of asthma in children under five, ” Prim Care Respir J, vol. 16, pp. 7–15, 2007.
[37] D. Firth, J. Glosup, and D.V. Hinkley, “Model Checking with nonparametric curves,” Biometrica, vol. 78(2), pp. 245-252, 1991.