International Journal of Engineering &Technology, 7 (3.12) (2018) 769 -772

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Comparative Study of Regression Techniques in the Estimation of UPDRS Score for Parkinson's disease

Santhi.B[1*], Harini Ram Prasad[2], Rohith Jayaraman[3]

**Corresponding Author Email: [1]shanthi@cse.sastra.edu, [2]hramprasad98@gmail.com, [3]iamrj19@gmail.com*

Abstract

Studies have shown that instances of Parkinson's disease have been on the rise over the past 30 years. A metric that measures the extremity of Parkinson's disease in a person is their Unified Parkinson's Disease Rating Scale (UPDRS) score. Thus, an algorithm that can predict the UPDRS score of a Parkinson's patient will be effective in determining the severity of the patient's condition. This paper aims to forecast a patient's UPDRS score by inferring patterns from historical figures and other independent parameter values that affect the patients' UPDRS score. Four regression techniques namely multilinear, ridge, robust and LASSO regression are being used to predict the UPDRS scores. This will be done using the R language and through the use of the MASS, glmnet packages.

Keywords: *UPDRS, Parkinson's disease, Robust Regression, Multilinear Regression, LASSO Regression, Ridge Regression, Shimmer, Jitter, Voice measures, motor UPDRS*

1. Introduction

Parkinson's disease is a degenerative disease which causes people to lose control of their nervous system, and as a result, their motor skills. The severity of Parkinson's disease is measured by a system known as UPDRS. Since a patient's UPDRS score informs us of the extremity of the disease's progression, it allows the patients to take the necessary measures to cope with the disease. As such, an algorithm which predicts the UPDRS score of a patient of Parkinson's disease would be of great use to doctors as well as patients. Our paper aims to predict the UPDRS scores of patients of Parkinson's disease based on previously existing records of patients' UPDRS scores by extrapolating regression patterns present in the historic patient data. This data includes previously measured total UPDRS scores, motor UPDRS scores, voice frequency, shimmer, jitter measurements and ratios.

The total UPDRS score will be the predicted parameter. This will be predicted on the basis of several data attributes mainly including various measures of Shimmer, Jitter and the motor UPDRS scores. The regressive techniques that will be used to determine the total UPDRS scores are: multilinear, ridge, robust, and LASSO regression.

2. Related Works

In [1], the paper concentrated on different machine learning techniques used to predict the extremity of Parkinson's disease from recordings of a patient's voice. It used regular regression techniques as well as time series modelling and finally, employed simpler models such as hidden Markov models. The three different types of regressive models used were LASSO, ridge and linear regression to select the best or most influential features. In addition, moving average models were also used in order to employ the time series portion of the data set, by calculating a moving average variable which used the selected features from LASSO regression. This model was able to predict the severity with a 2% accuracy.

The research paper [2] also focuses on predicting the severity of Parkinson's disease in a patient through a measure and analysis of the patient's speech patterns. Regression techniques based on the use of support vector machines were employed to do so. Polynomial, radial function basis and sigmoid kernel function were used to implement different forms of support vector regression (SVR). As such, regression was used to predict, analyse and evaluate the severity of Parkinson's disease, particularly in relation to speech analysis. The experimentation technique was able to predict the UPDRS score to a decent accuracy with an absolute mean error of 5.7.

In [3], the focus is on trying to propose a new and more efficient model for predicting the presence of Parkinson's disease at an early stage. They try to do this through the identification of dysphonia. The study observed patients suffering from Parkinson's disease and analysed the data obtained through the use of neural networks and Support Vector Machines (SVM). Logical regression and classification was also used in order to make the analysis easy. The paper applies the concepts of rotation forest and Haar wavelets for the first time to propose a new robust model for efficient diagnosis of Parkinson's disease. The average correctness of prediction of the disease using different methods via the proposed model was around 92%.

In [4], the research paper focuses on the reasons and possible indications for Parkinson's disease. It is an analytical paper which explores the different health conditions and markers that can be used in the prognosis of Parkinson's disease. In addition to exploring motor-related markers, they discuss non-motor markers including loss in olfaction, autonomic dysfunction, vision, depression and many other factors. In addition, the importance of being able to diagnose or predict Parkinson's was discussed.

3. Proposed Model

The four models used in this paper are briefly explained below:

Multilinear Regression:

The multilinear regression technique is a method where the parameter or attribute that must be predicted is assumed to vary linearly based on multiple parameters. Thus, this method estimates the value of the required attribute by linearly interpolating the training data, and accordingly giving an approximation of its value for the given inputs. The lm() function was used to accomplish this in R. A linear regression model can be represented as :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Here, Y is defined as the total UPDRS score and x_1, x_2, so on are the various parameters such as: Jitter percentage, Shimmer percentage and others. All the other terms in the equation are constants which are estimated from the training data.

Ridge Regression:

Ridge regression is often compared to the OLS model. The OLS model tries to reduce the difference between the actual values and predicted values to make itself more efficient. Ridge also has lesser effects if outliers are sent in which is why it is a more advantageous option for generalising while OLS fits the given training data set better. It is implemented in R using the glmnet package. The ridge method uses L2 regularisation where attributes are to be penalised if they assign extra weightage to any attribute. So basically if any value gives wrong weightage then we penalise it by a term that is equal to the sum of the square of the coefficients.
The most commonly used ridge parameters are:
Alpha - used for a concept known as elastic net mixing. For ridge regression, alpha is usually zero.
Lambda - The value of lambda determines the influence of the penalty term that is assigned to attributes that are accorded more weightage than needed. When lambda is zero, it is the same as conventional regression. However, when the value of lambda is large, the coefficients will tend to zero. The six parameters, apart from weighted price, were passed through ridge regression as they were the regression parameters. The optimal value of lambda was found and applied.

LASSO Regression:

LASSO regression, in theory, is very similar to ridge regression. However, where ridge uses an L2 permutation function, LASSO uses L1 regression. So, in LASSO, instead of the sum of the squares of the coefficients, the sum of the modulus of the coefficients is used. Thus, the penalty is determined by the L1 module. LASSO regression is a feature selection model that works more efficiently in scenarios in which it is necessary to avoid using certain terms. This means assigning to some of the coefficients a value of zero. Robust regression cannot do this and hence, is not used in such situations. In terms of implementation in R, again it is very similar to ridge, except that the value of the alpha parameter is 1.

Robust Regression:

Robust regression involves extrapolating or identifying patterns, based on the analysis of outliers in the training data set. This is a very good alternative to the OSI model. The values of the outliers are considered even though they might have previously been present. It is implemented using the rlm() function found in the MASS package of R.

4. Experiment and Results

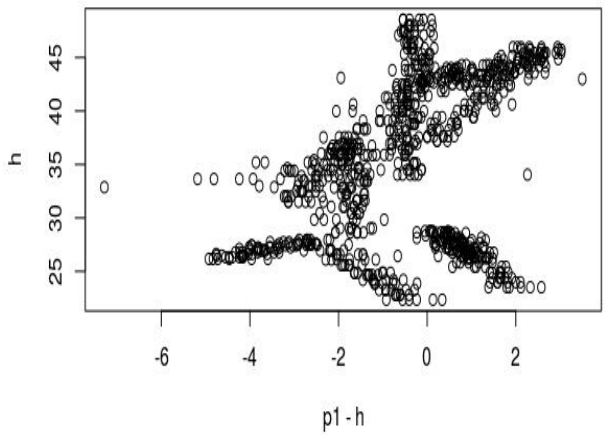
The Parkinson's dataset was taken from the UCI repository for data sets. It contained the UPDRS scores, Jitter, Shimmer and other relevant data fields. The training data consisted of 5000 rows and the last 876 instances were used as test data. Regressive predictive models as well as neural networks were used to give an accurate forecast of the test data. The total UPDRS score was predicted for the testing data and relative mean square errors were found. The values presented in the dataset had already been normalised and thus, no preprocessing was required.
The data set looks as follows:

subject	age	sex	test_time	motor_UPDRS	total_UPDRS	JitterPercent	JitterAbs	JitterRAP	JitterPPQ5	JitterDDP	Shimmer	ShimmerdB	ShimmerAPQ3	ShimmerAPQ5
1	72	0	5.6431	28.199	34.398	0.00662	0.00003380	0.00401	0.00317	0.01204	0.02565	0.230	0.01438	0.01309
1	72	0	12.6660	28.447	34.894	0.00300	0.00001680	0.00132	0.00150	0.00395	0.02024	0.179	0.00994	0.01072
1	72	0	19.6810	28.695	35.389	0.00481	0.00002462	0.00205	0.00208	0.00616	0.01675	0.181	0.00734	0.00844
1	72	0	25.6470	28.905	35.810	0.00528	0.00002657	0.00191	0.00264	0.00573	0.02309	0.327	0.01106	0.01265
1	72	0	33.6420	29.187	36.375	0.00335	0.00002014	0.00093	0.00130	0.00278	0.01703	0.176	0.00679	0.00929
1	72	0	40.6520	29.435	36.870	0.00353	0.00002290	0.00119	0.00159	0.00357	0.02227	0.214	0.01006	0.01337
1	72	0	47.6490	29.682	37.363	0.00422	0.00002404	0.00212	0.00221	0.00637	0.04352	0.445	0.02376	0.02621
1	72	0	54.6400	29.928	37.857	0.00476	0.00002471	0.00226	0.00259	0.00678	0.02191	0.212	0.00979	0.01462
1	72	0	61.6690	30.177	38.353	0.00432	0.00002854	0.00156	0.00207	0.00468	0.04296	0.371	0.01774	0.02134
1	72	0	68.6880	30.424	38.849	0.00496	0.00002702	0.00258	0.00253	0.00773	0.03610	0.310	0.02030	0.01970
1	72	0	75.6530	30.670	39.340	0.00465	0.00002553	0.00238	0.00260	0.00715	0.02132	0.188	0.01069	0.01214
1	72	0	82.6530	30.917	39.834	0.00537	0.00003216	0.00236	0.00278	0.00709	0.02377	0.282	0.01001	0.01375
1	72	0	89.6350	31.309	40.412	0.00524	0.00003287	0.00235	0.00251	0.00704	0.02493	0.240	0.01176	0.01395
1	72	0	96.6330	31.776	41.034	0.00354	0.00002388	0.00142	0.00150	0.00427	0.02107	0.171	0.00847	0.01040
1	72	0	103.6400	32.243	41.657	0.00530	0.00003181	0.00241	0.00231	0.00724	0.02791	0.291	0.01310	0.01260
1	72	0	110.6500	32.710	42.280	0.00456	0.00002908	0.00152	0.00194	0.00457	0.02878	0.264	0.01379	0.01494
1	72	0	117.6600	33.178	42.904	0.00693	0.00003930	0.00329	0.00285	0.00987	0.02810	0.274	0.01468	0.01430
1	72	0	124.6400	33.643	43.524	0.00652	0.00003783	0.00313	0.00311	0.00940	0.03011	0.320	0.01603	0.01733
1	72	0	131.6400	34.109	44.146	0.00571	0.00003711	0.00296	0.00293	0.00889	0.02522	0.223	0.01260	0.01466
1	72	0	139.6900	34.646	44.861	0.00372	0.00002221	0.00181	0.00195	0.00542	0.03230	0.288	0.01458	0.01732
1	72	0	145.6400	35.043	45.390	0.00285	0.00001646	0.00079	0.00109	0.00237	0.01524	0.133	0.00667	0.00682
1	72	0	152.6400	35.509	46.013	0.00629	0.00003574	0.00278	0.00293	0.00835	0.03791	0.338	0.01915	0.02174

ShimmerAPQ5	ShimmerAPQ11	ShimmerDDA	NHR	HNR	RPDE	DFA	PPE
0.01309	0.01662	0.04314	0.014290	21.640	0.41888	0.54842	0.160060
0.01072	0.01689	0.02982	0.011112	27.183	0.43493	0.56477	0.108100
0.00844	0.01458	0.02202	0.020220	23.047	0.46222	0.54405	0.210140
0.01265	0.01963	0.03317	0.027837	24.445	0.48730	0.57794	0.332770
0.00929	0.01819	0.02036	0.011625	26.126	0.47188	0.56122	0.193610
0.01337	0.02263	0.03019	0.009438	22.946	0.53949	0.57243	0.195000
0.02621	0.03488	0.07128	0.013260	22.506	0.49250	0.54779	0.175630
0.01462	0.01911	0.02937	0.027969	22.929	0.47712	0.54234	0.238440
0.02134	0.03451	0.05323	0.013381	22.078	0.51563	0.61864	0.200370
0.01970	0.02569	0.06089	0.018021	22.606	0.50032	0.58673	0.201170
0.01214	0.01844	0.03206	0.017443	25.672	0.49892	0.61068	0.173870
0.01375	0.02395	0.03003	0.017115	24.204	0.46686	0.57984	0.193900
0.01395	0.02019	0.03528	0.011876	22.203	0.56600	0.60571	0.209840
0.01040	0.01920	0.02540	0.015008	24.614	0.61348	0.60661	0.158810
0.01260	0.02069	0.03930	0.018093	23.533	0.51577	0.56790	0.214610
0.01494	0.02309	0.04138	0.020181	22.203	0.51806	0.56978	0.175080
0.01430	0.01952	0.04405	0.041980	20.878	0.52874	0.57711	0.349480
0.01733	0.02293	0.04810	0.031634	22.212	0.50991	0.61093	0.230480
0.01466	0.02145	0.03780	0.031546	23.129	0.52714	0.59220	0.182110
0.01732	0.02908	0.04373	0.010976	22.939	0.49687	0.57726	0.165670
0.00682	0.01299	0.01702	0.004652	25.181	0.42536	0.54735	0.169460
0.02174	0.03315	0.05745	0.043582	20.757	0.58088	0.56681	0.279240

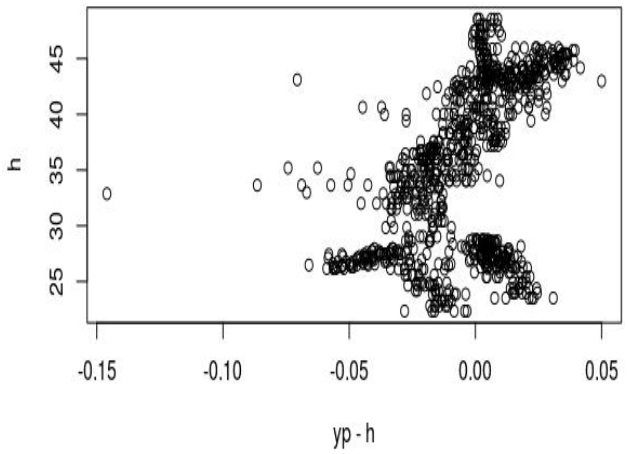
Multilinear Regression:

The multilinear regression model took the total UPDRS score of the patients as a function of all the remaining parameters or data fields in the table. The root mean square error (RMSE) for multilinear regression was 1.80.



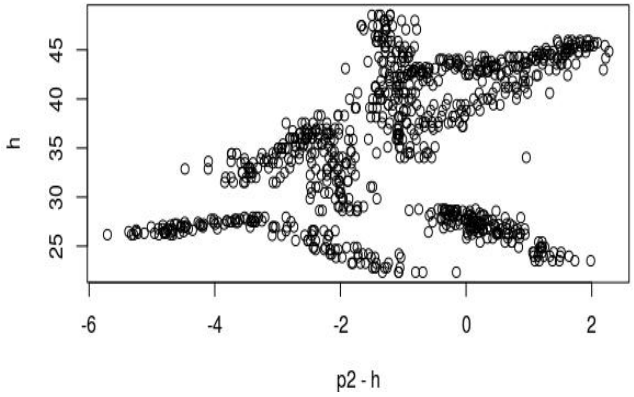
Ridge Regression:

The ridge regression technique requires the estimation for a value of lambda. In this case, lambda was taken to be the minimum or optimal value within the sequence 10^{-2} and 10^{3} in increments of 0.1. This is generally supposed to be the optimal range for finding the optimum value of lambda. The error calculated in ridge regression was 0.02 and was the least of the four regressive models, making it the most accurate predictive model for this case study.



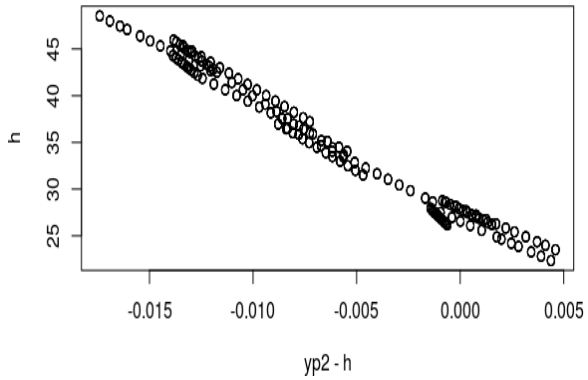
Robust regression:

Robust regression is an extended version of the linearly regressive model which again predicts the total UPDRS score, using the rest of the data fields as independent parameters. In robust regression, the influential observations are accordingly given penalties to help forecast the UPDRS score of the testing data. RMSE value was observed to be 2.00.



Least Absolute Selection and Shrinkage Operator (LASSO) Regression:

The LASSO model of regression takes a weightage factor equivalent to the magnitude of the coefficients. This model allows the elimination of certain coefficients which are reduced to 0 while larger penalties give smaller coefficients. The value of lambda, in this case, was taken to be 1. The resulting RMSE value was 0.15.



A snippet of the predicted values is shown below:

Actual value	MLR	Robust Regression	Ridge Regression	Lasso Regression
43.811	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
44.268	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
44.795	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
45.326	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
45.866	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
46.393	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
47.079	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
47.459	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
47.994	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
48.53	43.0256786939638	42.2483475490233	43.8056654376105	43.7979112170468
43.811	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
44.268	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
44.795	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
45.326	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
45.866	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
46.393	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
47.079	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
47.459	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
47.994	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
48.53	44.3272400064188	42.2483475490233	43.8056654376105	43.7979112170468
43.811	44.8319204459006	42.2483475490233	43.8056654376105	43.7979112170468
44.268	44.8319204459006	42.2483475490233	43.8056654376105	43.7979112170468

The error values can be summarised as follows:

Method	RMSE
MLR	1.80
Robust	2.00
Ridge	0.022
LASSO	0.15

5. Conclusion

In the above regressive models, multilinear regression had an RMSE value of 1.80, ridge regression had RMSE of 1.78, RMSE of robust regression was 2.00 and LASSO regression had an RMSE of 1.79. Amongst the error values calculated in the different regression techniques, ridge regression had the lowest error observed and robust regression had the highest RMSE value. As we can see graphically, ridge regression is the best because it has more values that are centred around zero error, which means it is the most accurate. For the given dataset, we may assume that the penalty for overweighted attributes works well with the penalty which is the sum of squares of coefficients. As such, in

this case, ridge regression becomes the most efficient and we may conclude that it works best as a predictive model.

References

[1] Nicolas Genain, Madeline Huberth, Roshan Vidyashankar, "Predicting Parkinson's Disease Severity from Patient Voice Features", http://roshanvid.com/stuff/parkinsons.pdf

[2] Meysam Asgari, Izhak Shafran, "Predicting Severity of Parkinson's Disease from Speech", https://pdfs.semanticscholar.org/65ed/35f9aaa9aa140555a794412c668b57659fff.pdf

[3] Indrajit Mandal, N.Sairam, "Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system" Volume 82, Issue 5, May 2013, Pages 359-377. https://ac.els-cdn.com/S1386505612001980/1-s2.0-S1386505612001980-main.pdf?_tid=spdf-c55764c2-848d-47d5-a63f-2e83269b82a4&acdnat=1519628353_c1580d56687b204397b8d285c0bba363

[4] R.B.Postuma, J.Montplaisir, "Predicting Parkinson's disease – why, when, and how?" Parkinsonism & Related Disorders Volume 15, Supplement 3, December 2009, Pages S105-S109, https://ac.els-cdn.com/S135380200970793X/1-s2.0-S135380200970793X-main.pdf?_tid=spdf-453f3fca-b586-4f1b-86ee-eb39821eb8a9&acdnat=1519628423_f2bcd4dfb70cc159dfb11f8b6c2a5651

[5] Gokul.S, Sivachitra.M, Vijayachitra.S, "Parkinson's Disease Prediction Using Machine Learning Approaches" Fifth International Conference on Advanced Computing (ICoAC) 2013,http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6921958

[6] A Tsanas, MA Little, PE McSharry, LO Ramig, "Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests" IEEE transactions on Biomedical Engineering, Vol. 57, No. 4, April 2010, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5339170

[7] Trevor Hastie, Junyang Qian, "Glmnet Vignette" June 26, 2014,https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

[8] Pauline Anderson, "Parkinson's Disease on the Rise?" MedScape March 6, 2018,https://www.medscape.com/viewarticle/865262

[9] Parkinson's foundation, "Understanding the Parkinson's disease. What is Parkinson's?", http://www.parkinson.org/understanding-parkinsons/what-is-parkinsons

[10] Selva Prabhakaran, "Robust regression", http://r-statistics.co/Robust-Regression-With-R.html

[11] University of Michigan, "Multiple Linear Regression",http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat401/Notes/401-multreg.pdf

[12] Stephanie, "LASSO Regression" October 12th 2017, http://www.statisticshowto.com/lasso-regression/

[13] Parkinson's UK, "The Unified Parkinson's Disease Rating Scale" (10 February 2016) https://www.parkinsons.org.uk/professionals/resources/unified-parkinsons-disease-rating-scale