ICT FOR HEALTH

Report n. **1**

"Linear Regression on Parkinson data"

Student: Gianluca Amprimo, s259455

Academic Year: 2019/2020

1. Introduction

1.1 Parkinson's disease and UPDRS

Parkinson's disease is a pathology for which is hard to identify the severity using an objective technique. The standard approach that is used by physicians is UPDRS, Unified Parkinson's Disease Rating Scale: the evaluation is done on the basis of a series of trials performed by patients, mainly involving some movements that, due to the illness, gradually become difficult for them (e.g.: tapping the other four fingers with the thumb, rising from a chair, walking a short distance). For each trial, a score is assigned and the sum of all the scores gives the "total UPDRS", that is used to tune the treatment, based on a medication called Levodopa. This kind of evaluation, however, requires a lot of time to be carried out and, because of the continuous progression of the illness, should be performed often. Moreover, it is not completely objective, because different doctors could disagree about the score of few points.

Therefore, it would be useful to identify an approach for computing total UPDRS that provides similar results, but that could be carried out by the patient himself, allowing to estimate the severity of the illness many times throughout the day and in a more "objective" form. This could help doctors in prescribing the correct quantities of Levodopa to alleviate the symptoms.

From literature, one approach that has been proposed is based on the evaluation of UPDRS according only to voice parameters, which could be automatically recorded by patients, considering that throat muscles are also affected by Parkinson's disease.

The goal of this paper is to evaluate the quality of UDPRS estimations performed by a linear regression model trained on this kind of data.

1.2 Linear regression

Linear regression is a mathematical model that tries to identify the trend of a certain feature y in the data (the regressand) according to other features, generally collected in a feature vector x (the regressor). In this case study, y represents the total UPDRS, that we want to estimate according to voice parameters in x, taken from the patient. The model is trained starting from a dataset containing information about previous patients, which is used to define the model equation:

$$y = \mathrm{X}\mathrm{w} + \upsilon$$

in which X is a matrix containing, in each row, the features of a patient, y is a column vector with the values of the regressand for each patient, w is the weight column vector that is used to linearly weight the different features, and υ is a measurement error. The goal of training a good linear regression model is to identify a w that, according to the y and X coming from the training dataset, minimizes the estimation error, so that each new prediction is as close as possible to the total UPDRS a doctor would assign to a patient. We can define the quantity to minimize as

$$\|y - Xw\|^2$$

This optimization problem can be solved using different approaches. In this paper, four different techniques will be compared:

- Linear Least Squares
- stochastic gradient algorithm with Adam optimizer
- conjugate gradient algorithm
- ridge regression.

For each approach, the distribution of the estimation error ($\mathrm{UPDRS}_{\mathrm{predicted}} - \mathrm{UPDRS}_{\mathrm{real}}$) for training, validation and testing and its mean, standard deviation, mean square value will be computed and

compared. Moreover, to provide an overall evaluation of the linear regression models obtained using the different algorithms, the coefficient of determination R^2 (only for the test data set) will be computed. This is defined as

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_d^2} = 1 - \frac{\sum_{n=1}^{N}[\hat{y}(n) - y(n)]^2}{\sum_{n=1}^{N}[y(n) - \bar{y}]^2}$$

with σ_e^2 the variance of the estimation error, σ_d^2 the variance of the data in the test set, $\hat{y}(n)$ the estimated value of total UPDRS for patient n and $\bar{y}$ the mean value of total UPDRS in the test set. Its value should be, in an accurate linear regression, as close as possible to 1, meaning that the ratio of the variances tends to 0, hence $\sigma_e^2 \ll \sigma_d^2$.

1.3 Dataset analysis

1.3.1 Dataset description

The dataset employed in this paper contains 5875 entries with 22 features and has been produced using data from 42 people with early stage of Parkinson's disease, recruited to a six-month trial of a telemonitoring device for remote symptoms progression monitoring. An entry is a voice recording performed by one of the 42 people, with around 200 recordings for each patient. The recordings were automatically captured in the patient's home. No missing values are present in the dataset, table 1 contains a summary of all the features.

Table 1: Features list and their characteristics

Feature name	**Type**	**Brief description**
subject#	numerical	Identifier for patient
age	numerical	Patient age
sex	categorical	Patient sex (0-male, 1-female)
test-time	numerical	Time since recruitment into the trial. The integer part is the number of days since recruitment.
motor-UPDRS	numerical	Clinician's motor UPDRS score, linearly interpolated
total-UPDRS	numerical	Clinician's total UPDRS score, linearly interpolated
Jitter (%)	numerical	Several measures of variation in fundamental frequency
Jitter (Abs)		
Jitter: RAP		
Jitter: PPQ5		
Jitter: DDP		
Shimmer	numerical	Several measures of variation in amplitude
Shimmer (dB)		
Shimmer: APQ3		
Shimmer: APQ5		
Shimmer: APQ11		
Shimmer: DDA		
NHR	numerical	Two measures of ratio of noise to tonal components in the voice
HNR		
RPDE	numerical	A nonlinear dynamical complexity measure
DFA	numerical	Signal fractal scaling exponent
PPE	numerical	A nonlinear measure of fundamental frequency variation

1.3.2 Data pre-processing

Before applying any algorithm, the dataset needs to be pre-processed and split into a training set, a validation set and a test set.
First of all, the feature "Subject#" has to be removed from the data because it is not relevant for the regression task and could bias the result, as well as "test-time" which is not of interest for this kind of analysis. "Motor-UPDRS", instead, is retained, in order to produce a more evident result (the value of "total-UPDRS" has some degree of correlation to "motor-UPDRS" because data scattered with respect to these two features are close to the plane main diagonal, as shown from scatter plot in figure 1), even if in theory it should not be considered.

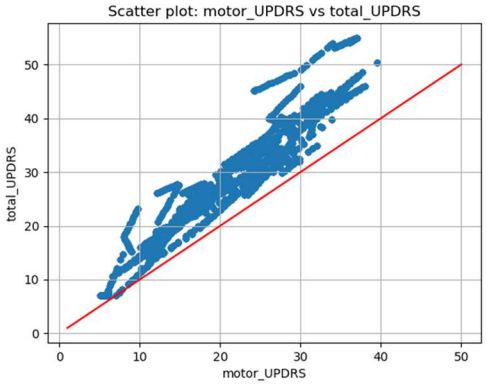

Figure 1: Scatter plot of motor-UPDRS vs total-UPDRS

Once the non-relevant attributes are removed, the dataset is shuffled and split into three sets: training, validation and test, with respectively 50%, 25% and 25% of the data.
The three sets are then normalized, using the mean and the standard deviation of the training set. The mean and standard deviations are evaluated only after splitting the data because test and validation should not affect the training phase, so they should not contribute to the parameters for the normalization of the data. In order to have the same normalization, however, also the test set and the validation set are normalized using mean and variance computed for training.
In the last step of pre-processing, "total-UPDRS" is extracted as the regressand y, whereas all the remaining 19 features are included in regressor vector x.

2. Linear regression methods and results

2.1 Linear Least Squares (LLS)

The Linear Least Square method computes vector w of linear regression using the formula

$$w = [X^T X]^{-1} X^T y$$

which applied to the training data ($X = X_{training}, y = y_{training}$) provides the result in figure 2.
The plot shows two strange peaks that outweigh all other values, one positive around +36 and one negative of -36, associated to features "Shimmer: DDA" and "Shimmer: APQ3". Scatter plot for these two attributes (fig. 3) shows that they are perfectly correlated (points are aligned along the plane main diagonal), therefore they are redundant and it would be reasonable to remove one of them: this operation makes the remaining feature more regular (fig. 4). However, even considering both, the regression model obtained produces good results, because the two contributions are similarly weighted but with opposite sign, hence they cancel each other. A similar discussion holds also for "Jitter: RAP" and "Jitter: DDP" that have the same behaviour (even if they produce smaller peaks).

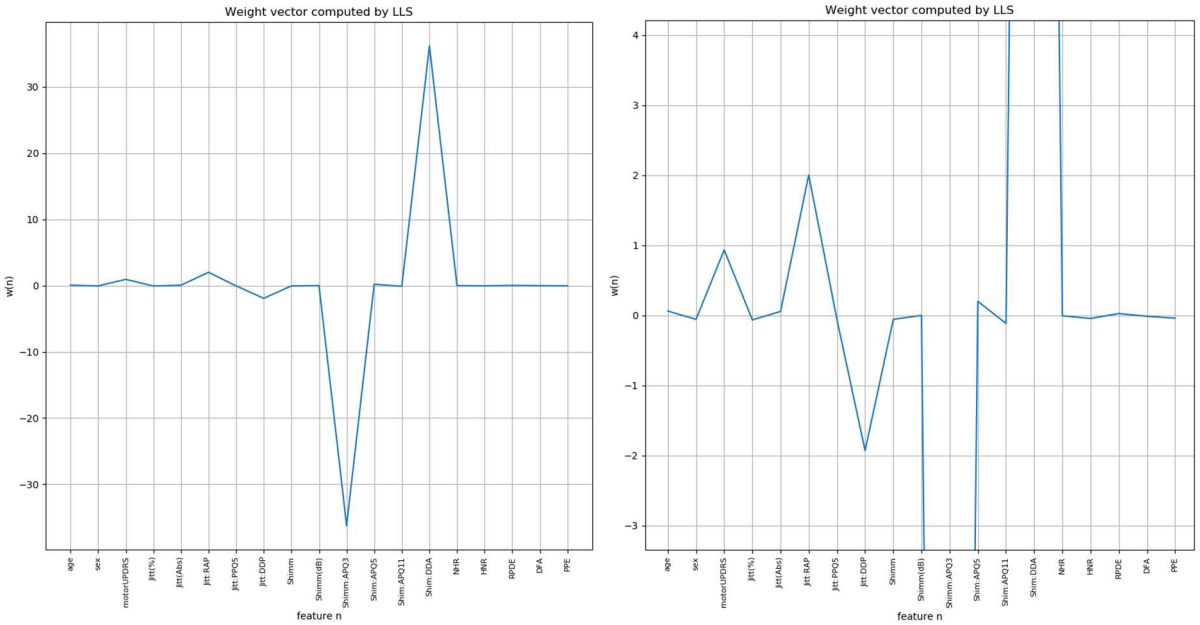

Figure 2: Weight vector computed by LLS- on the left scale (-40,-40), on the right scale (-3.5, 4.5)

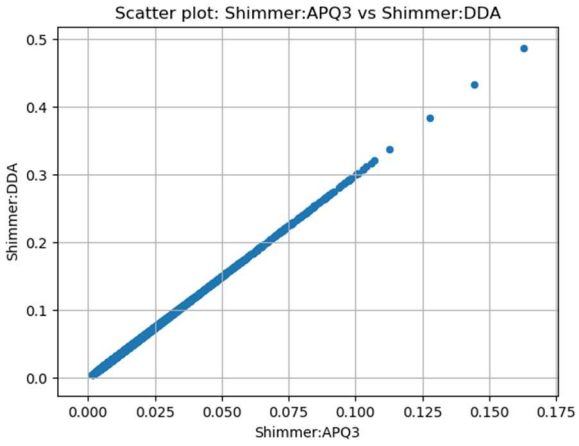

Figure 3: Scatter plot Shimmer:APQ3 vs Shimmer:DDA

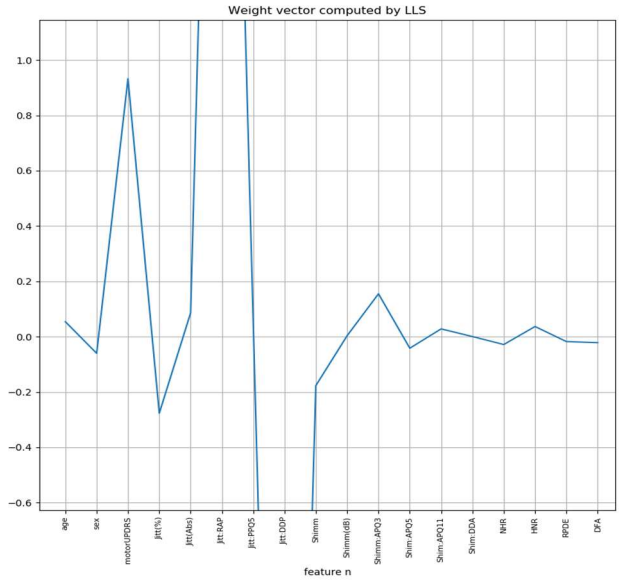

Figure 4: w for LLS removing Shimmer:APQ3

The histograms of the estimation error for training, testing and validation (fig. 5) all approximate a normal distribution, which essentially tends to 0 outside the (-5, 5) interval, apart from a small peak between (-10, -5). In general, the three distributions are slightly shifted to the right with respect to 0 and have their centre around 1, suggesting that the model tends more to overestimate by a few points. Instead, the small peak on the left highlights that large mistakes of 5-10 points frequently occur because the model has underestimated the score. This behaviour is also confirmed by the scatter plot of estimated vs true total UPDRS in figure 6. Moreover, the mentioned plot confirms that the prediction $\hat{y}$ in most of the cases is not so far from the real value y, because the points align around the plane main diagonal ($y \approx \hat{y}$).

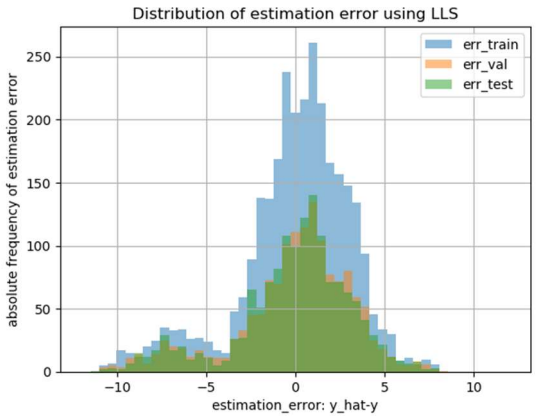

Figure 5: histograms of estimation error for LLS

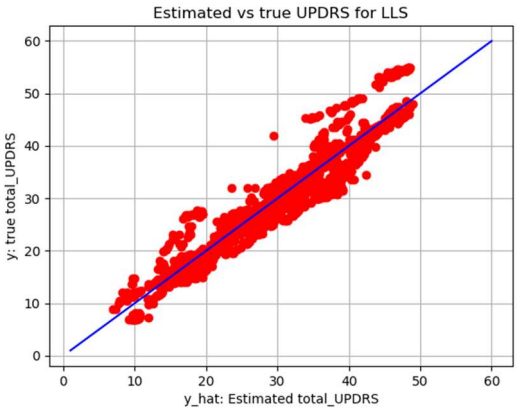

Figure 6: estimated vs true total UPDRS for LLS

2.2 Stochastic gradient descent with Adam method

Stochastic gradient descent algorithm with Adam optimizer requires two hyperparameters to tune, the number of iteration N_{it} and the learning coefficient γ. Moreover, also the values β_1, β_2 and ε for computing Adam's update at each step should be defined: in this case, they have been set by default to 0.9, 0.999 and 10^{-8}, as suggested in literature.

N_{it} and γ, instead, have been chosen with a "trial and error" approach, using the plot of MSE at each step of the algorithm. A valid combination is N_{it} =60,000 (around 21 epochs) and $\gamma=10^{-3}$, which produces a curve going close to 0 in 10,000 iterations (fig. 7). Nevertheless, N_{it} has been retained at 60,000, considering that this does not affect much the computational efficiency but produces slightly better result for R^2 and the other parameters of estimation error.

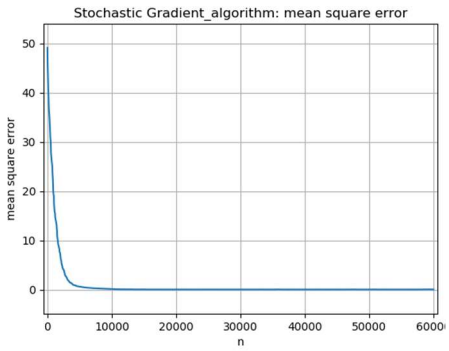

Figure 7: MSE of the estimation error vs number of iterations

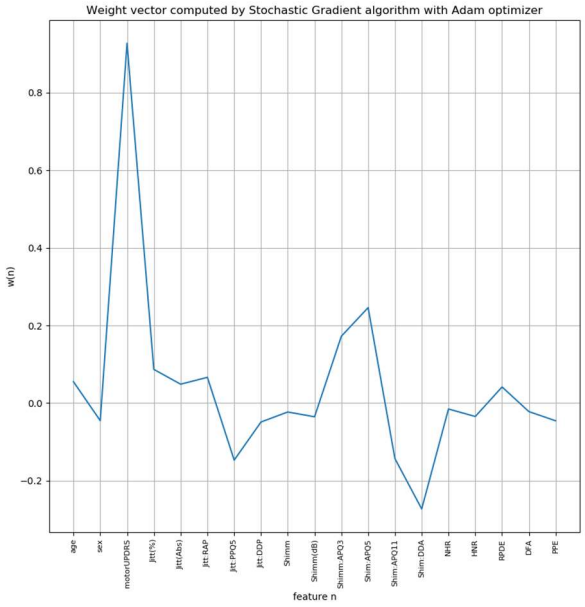

Figure 8: w computed by stochastic gradient algorithm

The resulting weight vector is shown in figure 8. All the weights are included in range (-0.25, 0.2) but for "motor-UPDRS" that provides the most significant contribution with a value around 0.9. The histograms of estimation error (fig. 9) for training, validation and test sets and the plot of estimated vs true total UPDRS (fig. 10) provide results analogous to the one discussed for LLS.

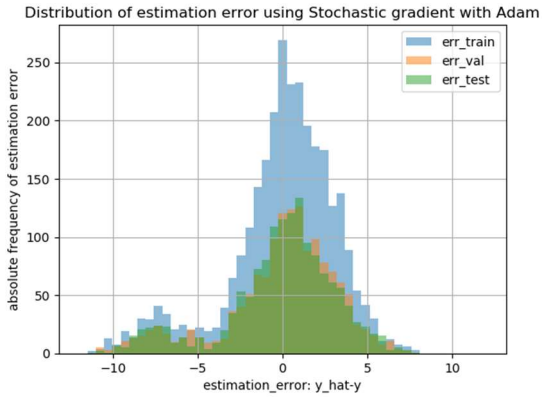

Figure 9: distribution of estimation error for stochastic gradient

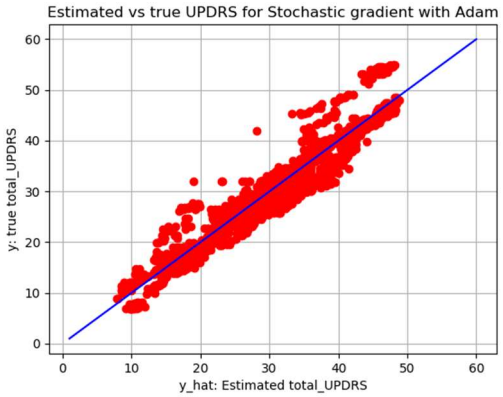

Figure 10: estimated vs true total UPDRS for stochastic gradient

2.2 Conjugate gradient

The conjugate gradient algorithm does not require to set any hyperparameters and it converges to the solution in a maximum number of steps equal to the number of features, 19 in this case.
The obtained w is shown in figure 11 and it is reasonably close to the result found for stochastic algorithm. The distribution of estimation error (fig. 12) in the different sets and the plot of estimated vs true total UPDRS (fig. 13) provide results analogous to the one discussed for previous methods.

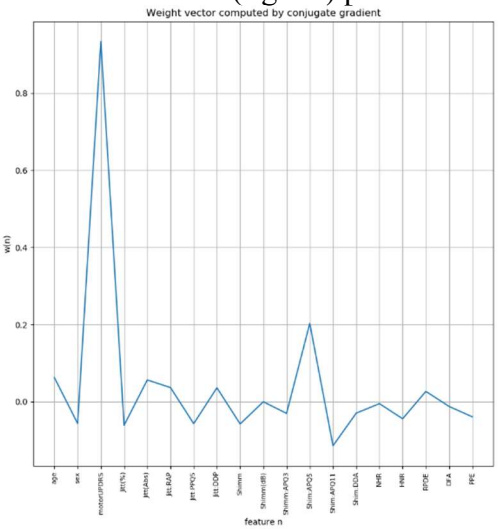

Figure 11: w computed by conjugate gradient

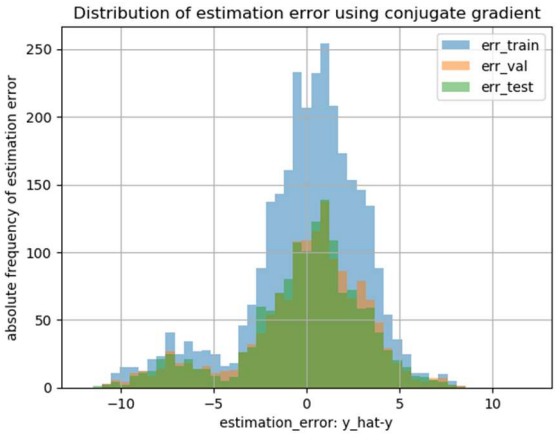

Figure 12: distribution of estimation error for conjugate gradient

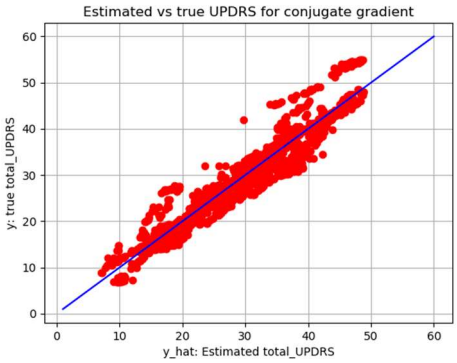

Figure 13: estimated vs true total UPDRS for conjugate gradient

2.2 Ridge regression

Ridge regression computes the value of the weight vector w using the formula

$$w = [X^T X + \lambda I]^{-1} X^T y$$

with I the identity matrix and λ an hyperparameter. In this paper, λ has been chosen plotting the MSE, on validation and training sets, of ridge regression models trained with values of λ between 0 and 100 (fig. 14). λ minimizing MSE for validation has been selected, which is λ=28.

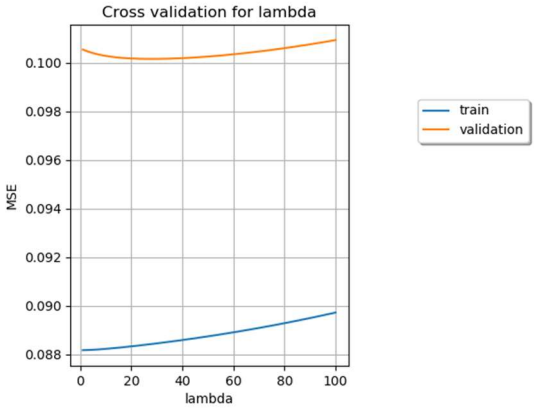

Figure 14: MSE (training/validation) of ridge regression vs λ

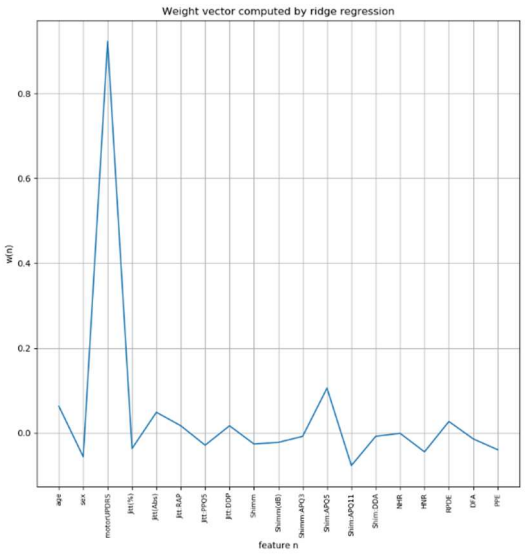

Figure 15: w computed by ridge regression

The computed weight vector for the chosen λ is shown in figure 15 and it is reasonably close to the results found for previous algorithms. The histograms of estimation error (fig. 16) in training, validation and test sets and the plot of estimated vs true total UPDRS (fig. 17) are analogous to the one discussed for previous methods.

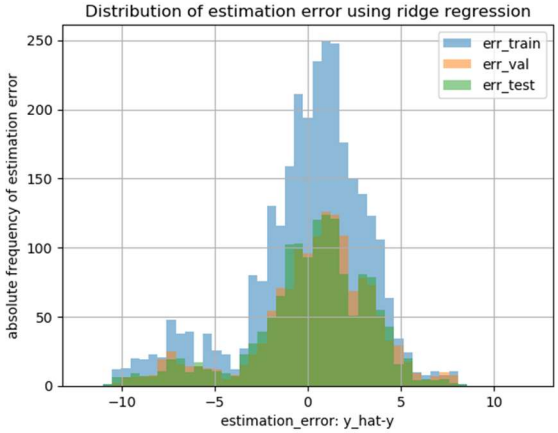

Figure 16: distribution of estimation error for ridge regression

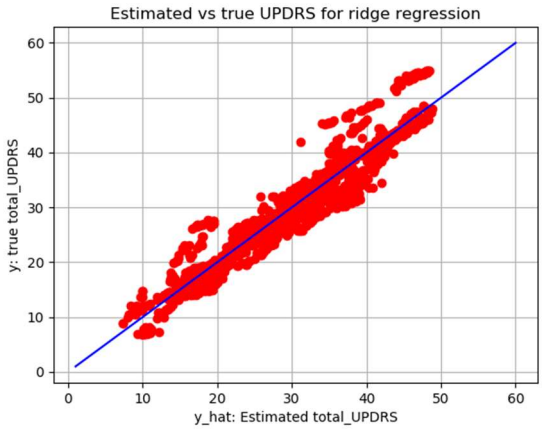

Figure 17: estimated vs true total UPDRS for ridge regression

3. Conclusions

Table 2 summarizes the parameters of estimation error (mean μ, standard deviation σ, mean square value MSE) obtained using the different linear regressions and the coefficient of determination R^2, computed only for test set.

Table 2- Parameters of estimation error for the different techniques over training, validation and test

Method	training			validation			test			
	μ	σ	MSE	μ	σ	MSE	μ	σ	MSE	R^2
LLS	-1e-13	3.168	10.038	-0.097	3.385	11.470	-0.138	3.315	11.011	0.906

stochastic	7e-14	3.216	10.343	-0.107	3.472	12.063	-0.166	3.381	11.457	0.902
conjugate	7e-14	3.169	10.044	-0.098	3.383	11.460	-0.139	3.315	11.009	0.906
ridge	7e-14	3.174	10.074	-0.100	3.376	11.410	-0.137	3.309	10.968	0.907

The trained models provide parameters for the estimation error that are essentially all equivalent and compliant with the histograms analysed in section 2 (fig. 5-9-12-16): all linear regressions have an error centred in 0 with a standard deviation around 3, that is reasonable considering that also doctors can disagree by few points about the total UPDRS score. Moreover, the MSE for training is comparable to MSE for validation and testing, suggesting that all the models do not suffer from overfitting, so they are general enough to provide a correct estimation for a new patient. The coefficient of determination R^2, that should be as close as possible to 1, is around 0.9 for all the methods, hence they provide good predictions, idea supported by the scatter plots of estimated vs true total UPDRS in section 2 (fig. 6, 10, 13, 17).
Figure 18 shows a comparison of the weight vectors evaluated using the different algorithms.

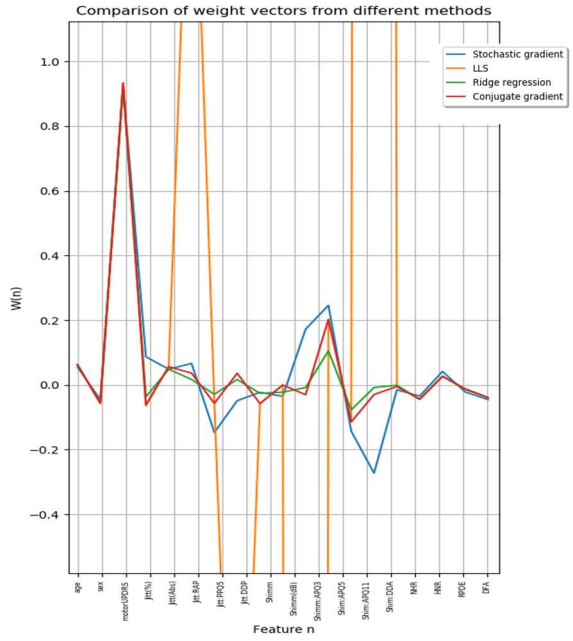

Figure 18: weight vectors computed by different methods

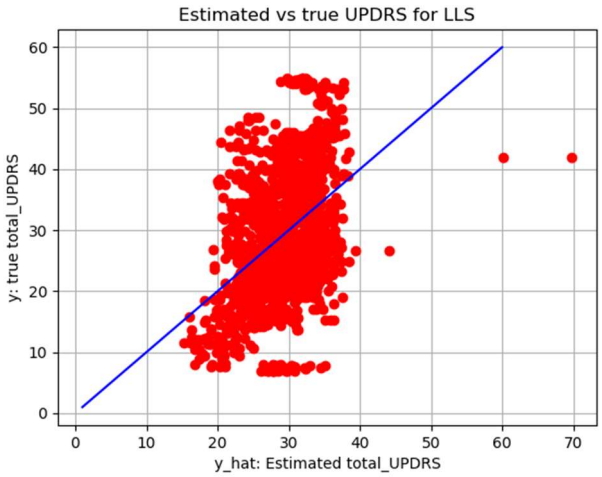

Figure 19: estimated vs true UPDRS for LLS without motor UPDRS

Without considering the strange peaks in LLS, already discussed in section 2.1, all the methods for regression tend to similar weights of the features, in particular for the first three ("age", "sex", "motor-UPDRS") and the last five ("NHR", "HNR", "RPDE", "DFA", "PPE"). Greater variability is related to weights of jitter and shimmer parameters; however, these variations do not seem to affect the results. It is likely that the minimized curve has a minimum which is a large plateau, hence the models work even if weight vectors are not all perfectly identical. In addition, it is evident from figure 18 that the feature with the most significant contribution to a correct UPDRS estimation is "motor-UPDRS", that has a weight of around 0.9, whereas all the other values are included in (-0.25, +0.25). Indeed, removing this parameter from the analysis makes the regression model completely unable to predict meaningful values (fig. 19), whatever the method for finding the optimal weight vector, with a very bad coefficient of determination R^2. We can conclude that a linear regression based only on these voice parameters is not enough to evaluate a trustworthy total UPDRS score, whereas if we consider also "motor-UPDRS", all the proposed methods are equally effective.