Analyzing a lower back pain data

A group project, done as a part of the Regression Techniques course
in M.Stat 1st Year (Fall 2020), Indian Statistical Institute, Kolkata
by Anik Burman (BS1730), Soham Das (BS1710), and Aditya Ghosh (BS1733),
under the guidance of Prof. Kiranmoy Das

Abstract

We analyze a dataset that contains the condition of the lower back pain of 310 persons (labelled Abnormal or Normal by trained physicians), collected along with some other physical spine details. Our goal is to identify which of those measurements might be important for predicting whether the pain in a person's lower back is Abnormal or Normal, and how to predict that using various regression models. Most medical studies carried out in the past were geared towards finding the importance of one of these predictors at a time. Here we consider all the predictors at once and assess their significance.

1 Introduction

Lower back pain can be caused by problems with any part of the complex, interconnected network of spinal muscles, nerves, bones, discs or tendons in the lumbar spine. While lower back pain is extremely common, the symptoms and severity of lower back pain vary greatly.

We analyze a dataset taken from kaggle, which has 310 observations, 13 attributes (12 numeric predictors, 1 binary class attribute). For each of the 310 individuals, their condition of pain in the lower back was labelled as Abnormal (210 many) or Normal (100 many) by some trained physicians. This will be our response variable, coding Abnormal as 1 and Normal as 0. The predictors in the dataset are described below.

- pelvic incidence (PI)
- pelvic tilt (PT)
- lumbar lordosis angle (LL)
- sacral slope (SS)
- pelvic radius (PR)
- cervical tilt (CT)
- pelvic slope (PS)
- Direct tilt (DT)
- thoracic slope (TS)
- degree-spondylolisthesis (DS)
- sacrum angle (SA)
- scoliosis slope (Sc.S)

Usage of abbreviations such as PI = Pelvic Incidence, PT = Pelvic Tilt, etc. are very common in the literature, which we follow here as well. The data do not contain any missing values. However, demographic information such as age, sex, body-weight etc. are absent.

Our goal is to identify whether the pain in a person's lower back is Abnormal or Normal using the collected physical spine details. Pelvic incidence (PI) is famous in the literature as an important factor for identifying Chronic Low Back Pain (see Ashraf et. al. [1], Ghobadifar [2], and the references therein). Spondylolisthesis is also known to be a major cause of lower back pain (see Rodts [3] and the references therein).

However, we could not find any literature where the above mentioned predictors (PI, DS etc.) are considered *all together*. Most studies are geared towards finding the importance of one of these predictors at a time. Hence one of our challenges was to consider all the predictors and assess their significance.

2 Multicollinearity?

Calculating the **multiple correlations for the predictors** reveals that there are evidences of multicollinearity. For instance, the **multiple correlation of pelvic incidence with the other predictors** turns out to be exactly 1. Investigating further, we found that the **multiple correlation of PI with PT and SS** is exactly 1, and the estimates for PT and SS are also both equal to 1. Actually we did not have the field knowledge. It turns out that it is indeed true that

$$\text{pelvic incidence (PI)} = \text{pelvic tilt (PT)} + \text{sacral slope (SS)}$$

The above relation follows by applying simple high school geometry in the following diagram.

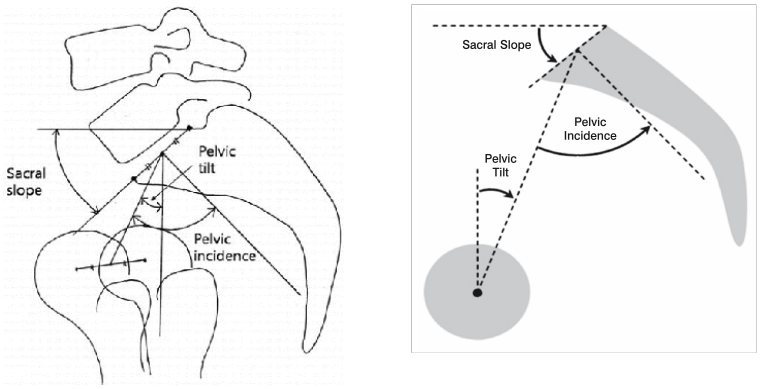

Figure 1: The relation PI = PT + SS

The multicollinearity rooted from this redundancy among these predictors. Now because of the relationship PI = PT + SS, we have to either drop one of these 3 predictors, or find some other way-out, for the sake of *identifiability*. We consider 3 subsets of the full data, by dropping the variables PI, PT and SS one-at-a-time, see Table 1. We also compare them with the case where only PI is retained and PT, SS both are dropped.

	PI	PT	SS
dat1	✗	✓	✓
dat2	✓	✗	✓
dat3	✓	✓	✗
dat4	✓	✗	✗

Table 1: Different subsets

Thinking of fitting regression models, note that considering dat1 essentially means setting the baseline that coefficient for PI is 0. Similar assumptions hold for the other dat's. Setting different baselines would make no change in the prediction when we consider the full model (e.g., a multiple logistic regression model with all the predictors). However, they will lead to different models when we consider model selection.

We already discussed that pelvic incidence (PI) is considered as an important predictor for detecting lower back pain (e.g., see [2]). In contrast, pelvic tilt (PT) and sacral slope (SS) are usually not given that much of attention. This is the reason why we are considering dat4.

Let us now check whether the multicollinearity is removed after removing the redundancy (by considering the 4 dat's). We find the variance inflation factors (VIFs) for the remaining predictors in each dat. A plot of these VIFs is shown in Figure 2. Observe that for each of the subsets (i.e., for each dat), none of the VIF's exceeds 10 (or even 5). Hence we can safely say that there is **no further sign of multicollinearity** being present among the remaining predictors.

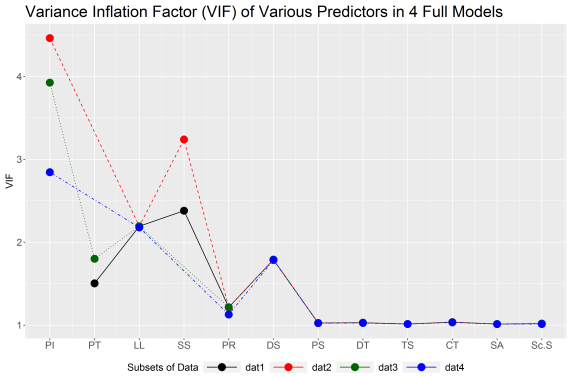

Figure 2: VIFs of the predictors after removing the redundancy

3 A first approach: Multiple Logistic Regression

First we apply the multiple logistic regression model with all the available predictors, on each of the datasets (dat1, dat2 etc.). We call these models as the *full models* for that particular subset of the data. Between dat1, dat2, and dat3, only the coefficients of PI, PT, SS will get changed for the change in baseline (which will also change the significance level of the coefficients), while the coefficients of the other predictors will be the same, and so will be the fitted values.

We illustrate the outputs of these fitted models using the *forest plot* in Figure 3. For illustration purpose, we used the model fitted on the scaled data (scaled so that each predictor has a unit variance), noticing that the inference about the significance of the coefficients remains unchanged even if we scale the predictors (since the z-statistics remain unchanged). Following are some observations from Figure 3:

1. For all the full models, each of the predictors PS, DT, TS, CT, SA, Sc.S has an estimated coefficient that is not significantly different from 0. So these predictors might not play an important role in the prediction.
2. We also see that **PT and SS are significant in all the full models they are present in**. However, their estimated coefficients have opposite signs. Since PI = PT + SS, it follows that when PT, SS both are absent (in Model 4), PI has a coefficient not significantly

different from 0. However, PI is significant in Models 2 or 3. Thus, **PI alone might not be significant, but becomes significant when PT or SS is present**.

These observations encourage us to try fitting a smaller model, which will be done next, using **best subset regression** or **penalized regression**.

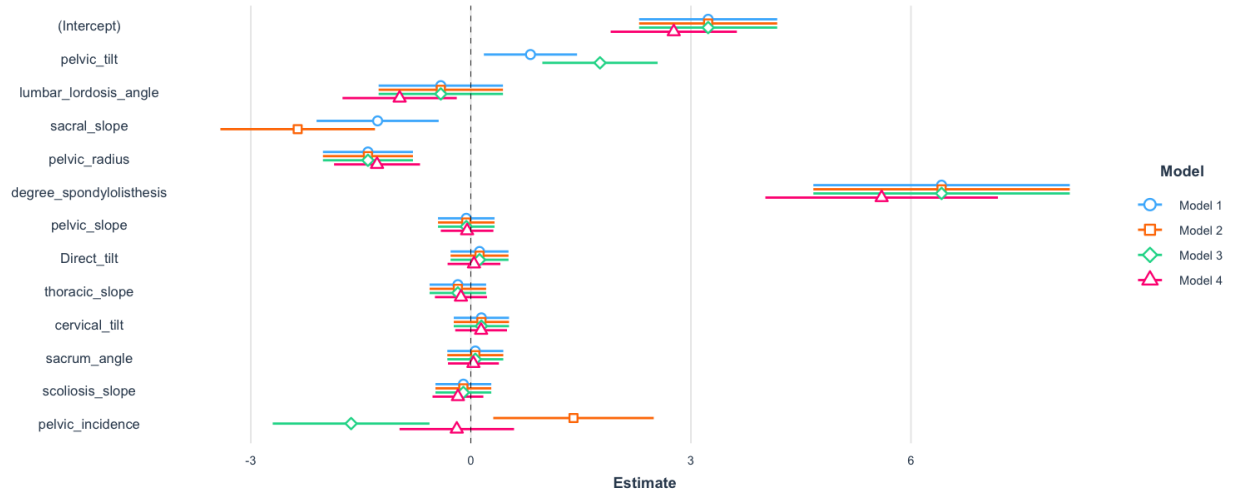

Figure 3: Forest plot for the fit of the full models for different dat's (Model k refers to the full model for dat$k$).

Before finishing this subsection, let us have a look at the standardized residuals of the full models. Since the residuals are the same for the full models 1, 2, and 3, in Figure 4 we only show the plot of standardized residuals against the fitted values for full model 1 and full model 4.

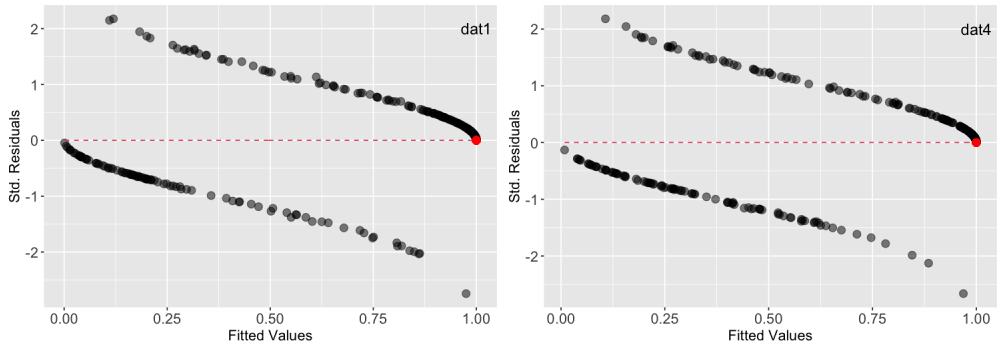

Figure 4: Plot of standardized residuals vs. fitted values for the two full models

We observe that a few standardized residuals (14 many for dat1, or 10 many for dat4), which are indicated by red in Figure 4, are almost equal to 0 (less than 10^{-3}). This exact-fit can be caused by either the presence of too many parameters, or due to some different reason, such as a complete or quasi-complete separation of the two classes. The latter possibility can be ruled out by noticing that the exact fit occurred for only a few observations, not for almost all the observations. To ensure that this problem of exact fit is not caused by the presence of too many parameters is another motivation for us to try fitting a smaller model (i.e., with less parameters).

4 Finding a smaller model

4.1 Best subset regression

We perform the best subset regression using the **forward method**, the **backward method**, and the **stepwise method** (starting both from the null model (only intercept) and the full model). We consider both the AIC and the BIC as our selection criteria. The results are shown below.

AIC	forward	backward	step-null	step-full
dat1	188.94	188.94	188.94	188.94
dat2	188.94	188.94	188.94	188.94
dat3	190.26	188.94	188.94	188.94
dat4	210.55	210.55	210.55	210.55

BIC	forward	backward	step-null	step-full
dat1	207.26	207.26	207.26	207.26
dat2	207.26	207.26	207.26	207.26
dat3	212.68	207.62	207.62	207.62
dat4	225.49	225.49	225.49	225.49

Table 2: AIC (resp. BIC) values of the subsets chosen by AIC (resp. BIC)

Observe that for dat3, using AIC (resp. BIC) as the criterion, the forward method chooses a subset having slightly higher AIC (resp. BIC) than the model selected in other subset selection methods (backward or stepwise), and the latter one also matches (in terms of AIC or BIC) with the best subset for dat1 and dat2. For dat1, these models are given by

$$\text{For AIC:} \quad \text{logit}(\mathbb{P}(\text{Abnormal})) = \beta_0 + \beta_{DS}x_{i,DS} + \beta_{PR}x_{i,PR} + \beta_{SS}x_{i,SS} + \beta_{PT}x_{i,PT} \qquad \text{(BS.A)}$$

$$\text{For BIC:} \quad \text{logit}(\mathbb{P}(\text{Abnormal})) = \beta_0 + \beta_{DS}x_{i,DS} + \beta_{PR}x_{i,PR} + \beta_{SS}x_{i,SS} \qquad \text{(BS.B)}$$

For dat2 or dat3, we have to adjust this model using the relation $PI = PT + SS$ (essentially changing the baseline for $\beta_{PI}, \beta_{PT}, \beta_{SS}$ whichever needed). In terms of both AIC and BIC, the best subset chosen for dat4 is found to be worse than the above two models. Thus, in context of the best subset regression, our best choice would be to use the two models mentioned above (BS.A and BS.B), fitted on dat1 (or, an equivalent model fitted on dat2, or dat3). These are indeed the best models (acc. to AIC or BIC) for dat1 (and also for dat2 and dat3).

Note that we can compare the models BS.A and BS.B because one is a subset of another. Of course, we cannot compare them by AIC or BIC because both are best accordingly as one of these criteria, but we can compare them using the residual deviances. The difference in the deviances of those two models gives a p-value of 0.020 (for the χ^2 test). So, the inclusion of effect of PT in the model does reduce the deviance significantly pelvic tilt (PT) once again.

Before ending this discussion, let us return to the issue of exact fit. In Figure 5 we show the plot of standardized residuals of logistic regression for these two models. We can see in Figure 5 that the exact fit for a few points is still present. But now we can understand that the problem is neither an issue of having too many parameters, nor a problem of complete/quasi-complete separation (because only a few points have the exact fit). So we decide to throw away those few points and do the analysis on the reduced data. Since those points had an exact fit, the deviances would not change after throwing them. Hence most of the analysis we did till now would remain

the same. In particular, the best subset chosen by AIC would remain the same, with the same value of AIC. On the other hand, each of the BIC values will increase by same amount (since n decreases), hence the selected model by BIC will be same again!

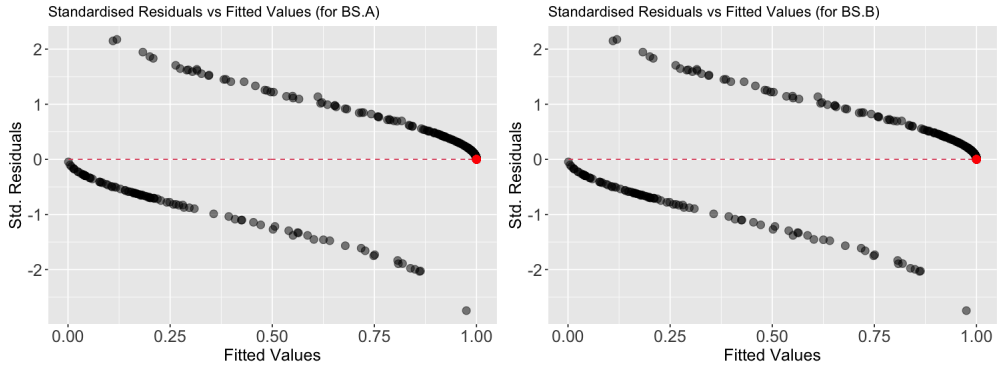

Figure 5: Plot of standardized residuals vs. fitted values for the two models BS.A (left panel) and BS.B (right panel). Note that the issue of exact fit is still present.

4.2 Penalized logistic regression

The main idea of penalized logistic regression is to introduce a penalty term in terms of the norm of the regression coefficients (excluding the intercept) in the log-likelihood of the usual multiple logistic regression model. For the penalized logistic regression we have to maximize the objective function

$$\frac{1}{N}\sum_{i=1}^{N}\left(y_i(\beta_0+x_i^\top\beta)-\log\left(1+e^{(\beta_0+x_i^\top\beta)}\right)\right)-\lambda\left((1-\alpha)\frac{\|\beta\|_2^2}{2}+\alpha\|\beta\|_1\right) \qquad (*)$$

which is to be maximized for $(\beta_0,\beta)\in\mathbb{R}^{p+1}$, where $\|\cdot\|_p$ is the ℓ_p norm, λ is a tuning parameter. The elastic-net penalty is controlled by α, which bridges the gap between Lasso ($\alpha=1$), and Ridge ($\alpha=0$). The tuning parameter λ controls the overall strength of the penalty.

The redundancy among the predictors caused a serious problem in the multiple logistic regression setup mainly because of the fact that the data matrix X is not full column rank, i.e., the matrix $X^\top X$ is actually a singular matrix. In contrast, it can be shown that while solving the optimization problem ($*$) we actually deal with $X^\top X+\lambda I$ instead of just $X^\top X$. For instance, the solution to the Ridge problem is given by

$$\widehat{\beta}_{\text{Ridge}}=(X^\top X+\lambda I)^{-1}X^\top y.$$

Hence the above problem of X being not full column rank gets resolved automatically, and there is no need to exclude any of PI, PT, SS beforehand. Moreover, for Lasso (or, for Elastic net with α close to 1), only a few coefficients are non-zero, so we also get the advantage of a subset selection method.

First, let us see the fit of the Ridge, Lasso and Elastic Net with $\alpha = 0.5$. The details on how an optimal λ is chosen are given below.

- In Figure 6, the estimated standard error bands for each estimated error rate are shown, computed by 10 fold cross-validation. The error measure taken here is the binomial deviance.
- We use the "one-standard-error" rule to choose the optimal λ – we pick the most parsimonious model within one standard error of the minimum. Such a rule acknowledges the fact that the trade-off curve is estimated with error, and hence takes a conservative approach.
- In each figure, the two selected λ's (rather $\log \lambda$'s) are indicated by the vertical dotted lines, where the left one corresponds to the minimum mean cross-validated error, while the right one corresponds to the value of λ that gives the most regularized model such that error is within one standard error of the minimum.

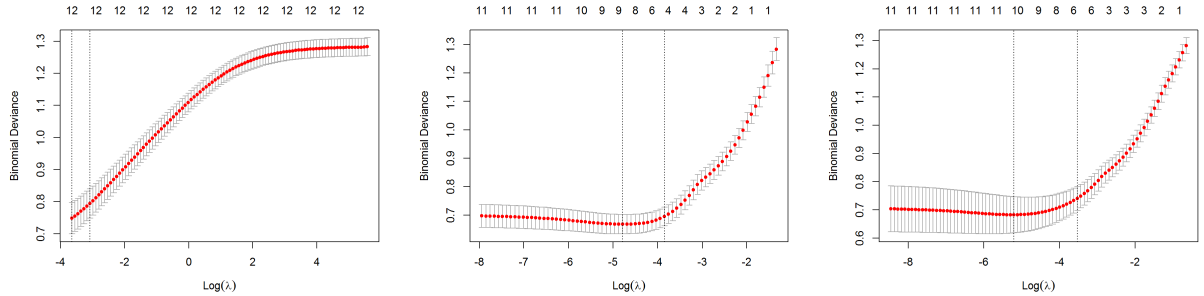

Figure 6: Choosing the optimal λ for Ridge (left panel), Lasso (middle panel), and Elastic Net with $\alpha = 0.5$ (right panel) by 10 fold cross-validation. Numbers in the top margin indicate the no. of fitted coefficients that are non-zero.

We see the with a smaller number of predictors, the optimal Lasso and Elastic net models chosen by the above method produce estimated error rates similar in magnitude to the chosen Ridge model. Figure 7 shows the 10 fold CV estimates of binomial deviances as a function of $\log \lambda$ for each of the models. Points on these curves that are circled indicates the model chosen by the one-standard-error rule. We see that among these three, Lasso performs the best.

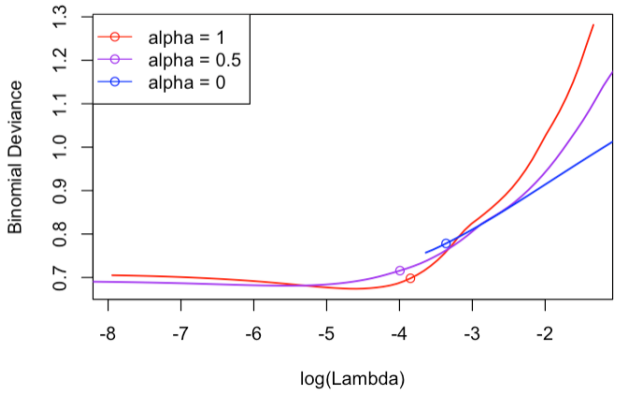

Figure 7: Comparing the deviances for elastic net penalty $\alpha = 1$ (Lasso), 0.5, and 0 (Ridge).

Figures similar to the Figure 7 were observed when the type measure in the cross validation is taken as the misclassification error rate or the area under the ROC curve. So we decide to take Lasso for the penalized regression model and compare it with the previously chosen models.

4.3 Final comparisons

To compare the fit of the Lasso model with the models chosen earlier, we can use the misclassification rate or the area under ROC curve (AUC). In Figure 8, the estimated standard error bands for each of these two cases are shown, computed by 10 fold cross-validation.

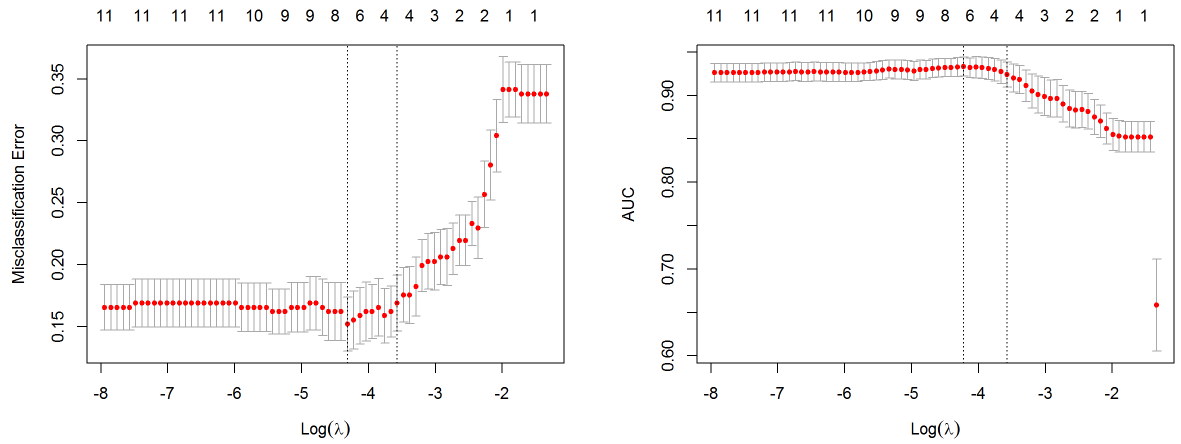

Figure 8: Optimally chosen Lasso models with type measure being misclassification error rate (left panel) or AUC (right panel).

The above two optimally chosen lasso models will be referred to as Lasso.class and Lasso.auc respectively. For both of them the predictors having non-zero fitted coefficients turn out to same as the predictors chosen in BS.A model (DS, PR, SS, and PT), the only difference being that the intercept and the fitted coefficients are smaller in magnitude for the lasso models. Table 3 records the fitted coefficients for all the 'optimal' models we discussed so far, along with the full models for dat 1 and dat4.

	Full1	Full4	BS.A	BS.B	Lasso.class	Lasso.auc
Intercept	15.182	13.961	15.455	18.051	5.650	6.168
PI		$*$			$*$	$*$
PT	0.081		0.066		0.021	0.023
SS	-0.095		-0.115	-0.117	-0.020	-0.026
DS	0.171	0.149	0.165	0.165	0.072	0.076
PR	-0.105	-0.096	-0.109	-0.121	-0.046	-0.049
LL	$*$	-0.052			$*$	$*$

Table 3: Fitted coefficients of different predictors for the models mentioned as the column names. A blank square indicates the absence of that particular predictor in the model, and $*$ indicates that it is not significant.

We now compare the previously chosen models with Lasso.class using their misclassification error rates and with Lasso.auc using their area under ROC curves (AUC). Table 4 shows the comparison using percentages of misclassification and Table 5 shows the comparison using AUC.

	Full1	Full4	BS.A	BS.B	Lasso.class
% Misclassification	18.4	16.6	18.1	16.8	16.5
# significant predictors	4	3	4	3	4

Table 4: Comparison using percentage of misclassification

	Full1	Full4	BS.A	BS.B	Lasso.auc
AUC	0.91	0.92	0.93	0.91	0.91
# significant predictors	4	3	4	3	4

Table 5: Comparison using area under ROC curve (AUC)

In terms of misclassification rates, the Lasso.class model performs equally well with the full model 4 (multiple logistic regression model for the dat4) and the BS.B model (best subset chosen by BIC); while all the models produce more or less similar values for the AUC. Thus, even though the Lasso models do tend to give the best error rates, those do not seem to be significantly better than the ones we get from the logistic regression model(s) with the best subset(s).

4.4 Other approaches

1. Performing linear and quadratic discriminant analyses (LDA and QDA respectively) by considering a randomly chosen half of the data as our training data and remaining half as the test data (and repeating this several times), we found that these methods give misclassification rates similar to the ones in Table 4, see the appendix for more details.

2. We also tried to fit a principal components regression (PCR) model. It was observed that for explaining at least 90% of the total variability one has to take at least 5 principal components (while doing PCA on the unscaled data). We fitted Lasso models with these principal components as the predictors and observed that this approach does give optimal models with similar error estimates, but requires more predictors than the Lasso models discussed earlier. This is why we decided to omit those from the report.

5 Conclusions and future scopes

Following points summarize the findings of our data analysis.

- The models chosen by the best subset regression with AIC has the predictors degree of spondylolisthesis (DS), pelvic radius (PR), sacral slope (SS) and pelvic tilt (PT). Exactly these predictors re-appear when we fit the Lasso (for all the three type measures).

- The multiple logistic regression model for dat4 and the BS.B model (best subset chosen by BIC) are on par with the Lasso model having type measure 'class', which tend to give the smallest percentages of misclassification among the five models discussed here.
- All the five models gave more or less similar AUC values. In particular, Lasso does not produce any better results than the logistic regression model(s) with the best subset(s).
- Pelvic incidence (PI) alone is not significant when PT, SS both are absent, but becomes significant when either of them are included.
- We observe from Table 3 that in all the optimally chosen regression models, PI is either absent, or does not have a significant contribution.

It is therefore suggested that **instead of using pelvic incidence alone, one should also take pelvic tilt and sacral slope into consideration**. The multiple logistic regression model *with or without a penalty* perform equally well here – adding a penalty does not give any added advantage except for tackling the redundancy in a proper manner.

Following are some future scopes of the project.

- We think that lots of work remain on understanding the relative importance of the measurements for detecting whether a person's lower back pain is abnormal.
- Demographic variables such as age, sex, height, weight etc. were not available in our data. These variables might have some important effect on detecting abnormal lower back pain.
- Field knowledge (e.g., measurement costs) might shed more light on which model to choose.

References

[1] Ashraf, A., Farahangiz, S., Jahromi, B. P., Setayeshpour, N., Naseri, M., & Nasseri, A. (2014). Correlation between radiologic sign of lumbar lordosis and functional status in patients with chronic mechanical low back pain. *Asian spine journal*, 8(5), 565.
Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4206805/

[2] Ghobadifar, M. A. (2015) Pelvic Incidence as a Determinant for Chronic Low Back Pain: Few Comments. *Asian spine journal*, 9(1), 149-150.
Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4330212/

[3] Rodts M. (2019) Spondylolisthesis: Back Condition and Treatment (online article).
Link: https://www.spineuniverse.com/conditions/spondylolisthesis/spondylolisthesis-back-condition-treatment

The codes used in this project are available at: ghoshadi.github.io/lower-back-pain.html.

Appendix

Fitted coefficients for penalized regression models

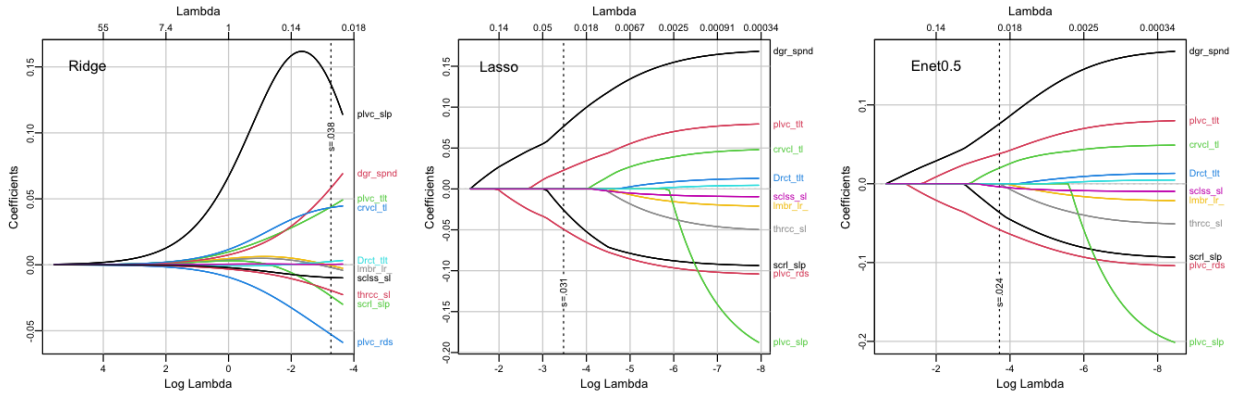

Figure 9: Fitted coefficients for Ridge (left), Lasso (middle) and Elastic Net with $\alpha = 0.5$ (right). Optimally chosen λ's (chosen by 10 fold cross-validation) correspond to the dotted lines.

Results of LDA and QDA

To perform the linear and quadratic discriminant analyses (LDA and QDA respectively), we consider half of the data as our training data and remaining half as the test data. Repeating this 20 times, we report the average percentages of misclassification in Table 6. Compare these with the misclassification rates in Table 4.

	dat1		dat2		dat3		dat4	
	Train	Test	Train	Test	Train	Test	Train	Test
LDA	16.23%	19.42%	16.06%	19.03%	15.45%	19.81%	18.61%	22.71%
QDA	10.97%	20.42%	11.45%	20.77%	11.45%	20.45%	14.35%	22.03%

Table 6: Percentages of misclassification for LDA and QDA for different subsets of the data