Final Project

Samira Zarandioon and Bradley Robinson

8/10/2018

Introduction

This data analysis report is based on a data set that is collected by the Global Longitudinal Study of Osteoporosis in Women (GLOW). According the official website of the research team (http://www.outcomes-umassmed.org/glow/), the goal of this research was to improve understanding of the risk and prevention of osteoporosis-related fractures among female residents of 10 countries who were 55 years of age and older. GLOW enrolled over 60,000 women through over 700 physicians in 10 countries, and conducted annual follow-up for up to 5 years through annual patient questionnaires. In this study, we explore how to better predict, using the given data, the likelihood that a female with osteoporosis will have a bone fracture given a number of predictors.

Data Description

The dataset used in this analysis is called GLOW500, which contains 500 observations from six sites in the united states. The target (outcome/response) variable FRACTURE indicates whether the subject had any fracture in the first year of followup. It includes 13 selected potential risk factors for fracture. According to the source of data, given the fact that fracture rate was only 4% these 500 observations are over sampled from 21,000 original observation.

Description of each field in the data set is as follows:

Explanatory Variables:

```
SUB_ID: Id code for each subject (Categorical)
SITE_ID: Id code for the study site (Categorical)
PHY_ID: Id code for the Physician (Categorical)
PRIORFRAC: Does the subject have history of prior fracture? (1 => Yes, 0 => No) (Bineary)
AGE: Age at enrolment (Numeric)
WEIGHT: Weight at enrollment (Numeric)
HEIGHT: Height at enrollment (Numeric)
BMI: Body mass index (Numeric)
PREMENO: Did the subjet had Menopause before age 45? (1 => Yes, 0 => No) (Bineary)
MOMFRAC: Has subject's mother ever had hip fracture? (1 => Yes, 0 => No) (Bineary)
ARMASSIST: Does the subject need arms to stand form a chare? (1 => Yes, 0 => No) (Bineary)
SMOKE: Is the subject a Smoker? (1 => Yes, 0 => No) (Bineary)
RATERISK: Subject's self-reported risk of fractur (1 => Less than others of the same age, 2 => Same as
FRACSCORE: Fracture risk score (Computed based on AGE, WEIGHT, ARMASSIST, SMOKE, MOMFRAC, PRIORFRAC)  (

Response Variable:

FRACTURE: Did subject have any fracture in first year? (1 => Yes, 0 => No) (Bineary)
```

A summary a of dataset that illustrates distribution of each filed is presented below:

```
##      SUB_ID         SITE_ID          PHY_ID         PRIORFRAC
##  Min.   :  1.0   Min.   :1.000   Min.   :  1.00   Min.   :0.000
##  1st Qu.:125.8   1st Qu.:2.000   1st Qu.: 57.75   1st Qu.:0.000
##  Median :250.5   Median :3.000   Median :182.50   Median :0.000
##  Mean   :250.5   Mean   :3.436   Mean   :178.55   Mean   :0.252
```

```
##  3rd Qu.:375.2   3rd Qu.:5.000   3rd Qu.:298.00   3rd Qu.:1.000
##  Max.   :500.0   Max.   :6.000   Max.   :325.00   Max.   :1.000
##       AGE             WEIGHT           HEIGHT           BMI
##  Min.   :55.00   Min.   : 39.90   Min.   :134.0   Min.   :14.88
##  1st Qu.:61.00   1st Qu.: 59.90   1st Qu.:157.0   1st Qu.:23.27
##  Median :67.00   Median : 68.00   Median :161.5   Median :26.42
##  Mean   :68.56   Mean   : 71.82   Mean   :161.4   Mean   :27.55
##  3rd Qu.:76.00   3rd Qu.: 81.30   3rd Qu.:165.0   3rd Qu.:30.79
##  Max.   :90.00   Max.   :127.00   Max.   :199.0   Max.   :49.08
##     PREMENO         MOMFRAC       ARMASSIST         SMOKE
##  Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   :0.00
##  1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.00
##  Median :0.000   Median :0.00   Median :0.000   Median :0.00
##  Mean   :0.194   Mean   :0.13   Mean   :0.376   Mean   :0.07
##  3rd Qu.:0.000   3rd Qu.:0.00   3rd Qu.:1.000   3rd Qu.:0.00
##  Max.   :1.000   Max.   :1.00   Max.   :1.000   Max.   :1.00
##     RATERISK      FRACSCORE         FRACTURE
##  Min.   :1.00   Min.   : 0.000   Min.   :0.00
##  1st Qu.:1.00   1st Qu.: 2.000   1st Qu.:0.00
##  Median :2.00   Median : 3.000   Median :0.00
##  Mean   :1.96   Mean   : 3.698   Mean   :0.25
##  3rd Qu.:3.00   3rd Qu.: 5.000   3rd Qu.:0.25
##  Max.   :3.00   Max.   :11.000   Max.   :1.00
```

Exploratory Analysis

Figure 1 shows a graphical representation of correlations between all the fields of glow500 datasets. This visualization gives us a very good understanding of how the explanatory variables are related to each other and the response variable. For example, it shows high correlation between PHY_ID and SITE_ID. Also, from this figure we can easily see that BMI is highly correlated to WEIGHT, which is driven by the definition of the BMI. Moreover, it shows that FRACSCORE is highly correlated with AGE, PRIORFRAC, etc., which are used to compute the risk. Correlated variables indicate some level of redundancy in the explanatory variables, and warrant feature selection to prevent overfitting.

Correlation of explanatory variables with the target variable FRACTURE is quite informative. The plot depicts that PRIORFRAC, AGE, ARMASSIST, and MOMFRAC have relatively high positive correlation to the target variable. However, HEIGHT has negative correlation to the target variable.

Performing Principal Component Analysis (PCA), against the numeric explanatory variables also gives us some insight on redundancy and importance of explanatory variables. As the Scree Plot inf figure 2 illustrates, only three components can explain over 99% of variance, which indicates some redundancy in these variables that can be justified due to correlation between these variables.

Addressing Objective 1:

We used step-wise parameter selection method to training a logistic regression model using maximum likelihood method. Parameters of trained model and their p-values are presented in figure 3

The corresponding regression formula is as follows:

$Logit(P_{Fracture}) = log(OddsRatioOfFracture) = log(P_{Fracture}/(1 - P_{Fracture})) = 3.2477 - 0.0378 * HEIGHT + 0.4072 * RATERISK + 0.2244 * FRACSCORE$

In other words:

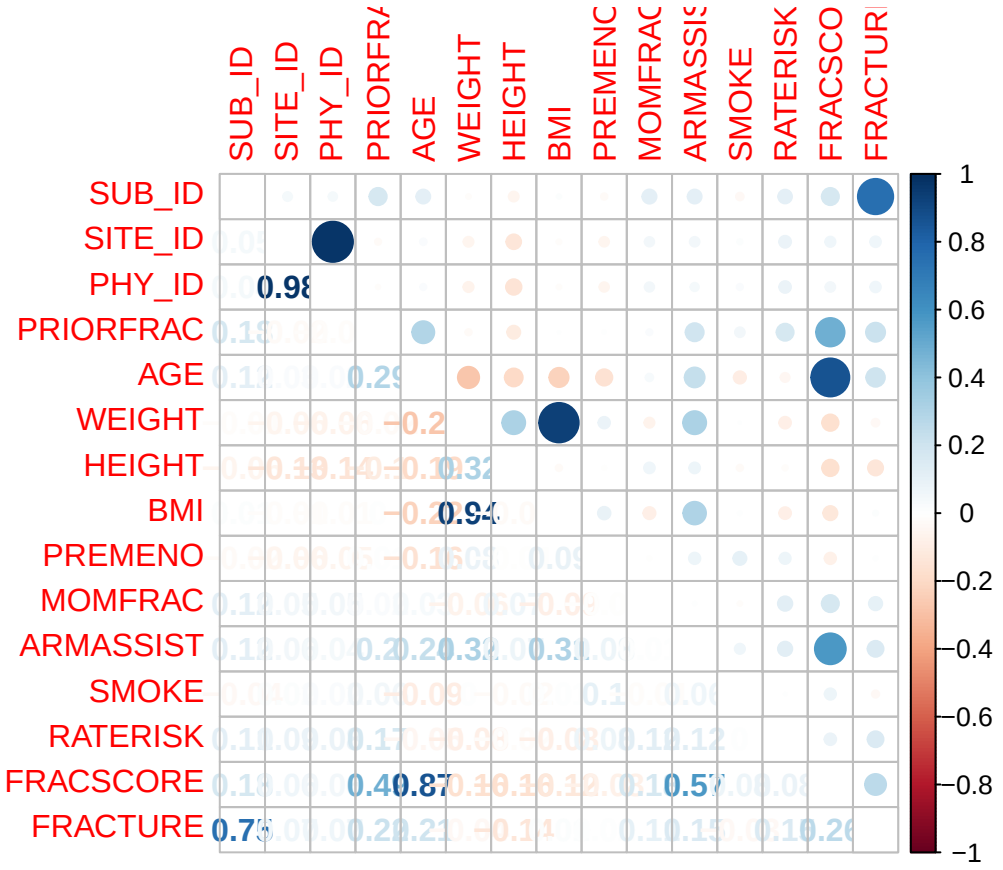

Figure 1: Correlation Matrix

Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
AGE	-.494722	0.467421	0.152466	0.716546	-.009160
WEIGHT	0.527303	0.465788	0.088410	0.032402	-.704363
HEIGHT	0.234577	-.081961	0.938232	0.018856	0.240042
BMI	0.474103	0.516152	-.245337	0.051374	0.667821
FRACSCORE	-.444298	0.539841	0.168723	-.694635	0.013601

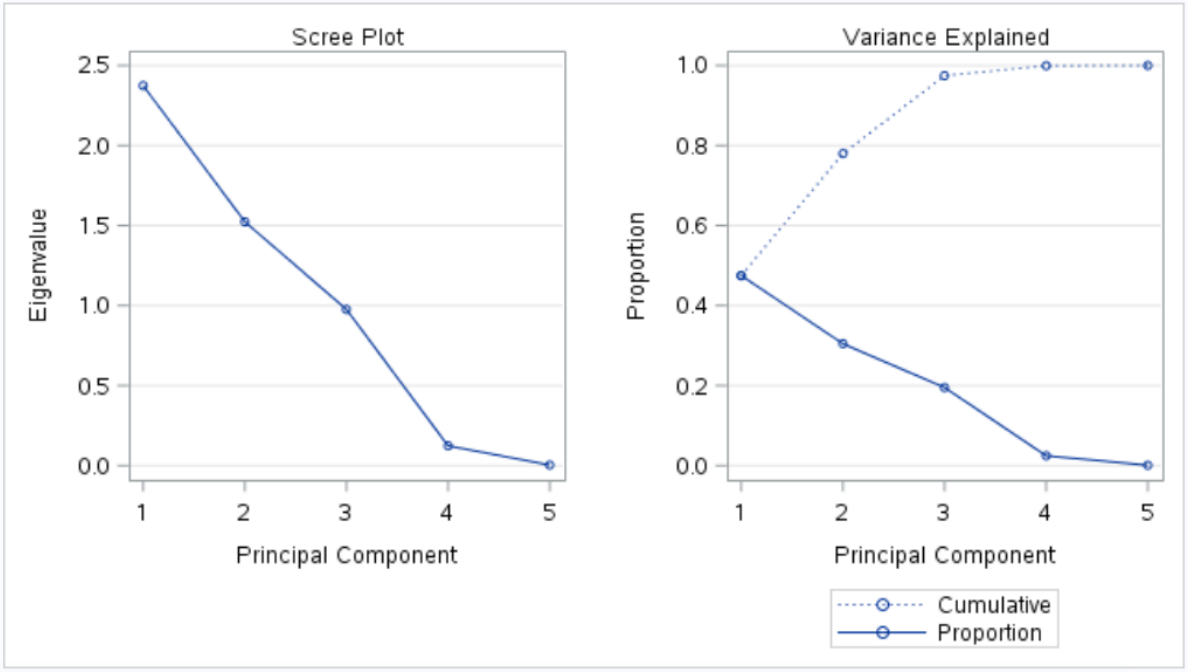

Figure 2: Principal Component Analysis On Numerical Fields

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.2477	2.8517	1.2970	0.2548
HEIGHT	1	-0.0378	0.0175	4.6608	0.0309
RATERISK	1	0.4072	0.1392	8.5576	0.0034
FRACSCORE	1	0.2244	0.0440	25.9729	<.0001

Figure 3:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	564.335	522.085
SC	568.550	538.943
-2 Log L	562.335	514.085

Figure 4: Model fit statistics

$$P_{Fracture} = Softmax(3.2477 - 0.0378 * HEIGHT + 0.4072 * RATERISK + 0.2244 * FRACSCORE)$$

Model Selection

We found that step-wise parameter selection finds the simplest and most efficient logistic regression model. In next sections we compare this model with other alternatives.

Table **??** shows how the three variables included in the model are selected and their corresponding p-value at each step:

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	FRACSCORE		1	1	34.9747		<.0001
2	RATERISK		1	2	8.7541		0.0031
3	HEIGHT		1	3	4.7043		0.0301

Model evaluation:

Model fit statistics for the selected model is shown in figure 4 deviance = 514.085.

These error metrics are based on the training dataset, so they only show how well the model is able to fit the data. But it does not show how well the model is able to generalize. We evaluated the same model using corss validation and produced the ROC curve in figure 5 that shows the specificity and sensitivity of the model and it's capability to generalize. The Area Under the Curve (AUC) is 0.6927, which is reasonabily good and signficicantlty better then arandom model (AUC = 0.5). In future sections we compare this model aginst other competing models.

Checking Assumptions

According to the Cook's D, it appears that within the three variables included in the model there are unlikely any values that are too influential (none are greater than 0.5). In addition, while there are no values that have large leverage in the training data and all are clustered close to each other. This makes it simpler to use all three variables without the concern that any one point would negatively affect the predictive value of the model.

Observing the lack of fit, we find that there is not enough evidence to reject the null hypothesis ($p > 0.05$) of the test that the data fits the model well. It appears that our model, using the three variables, fits the data well and that there are little concerns about outliers affecting the model output.

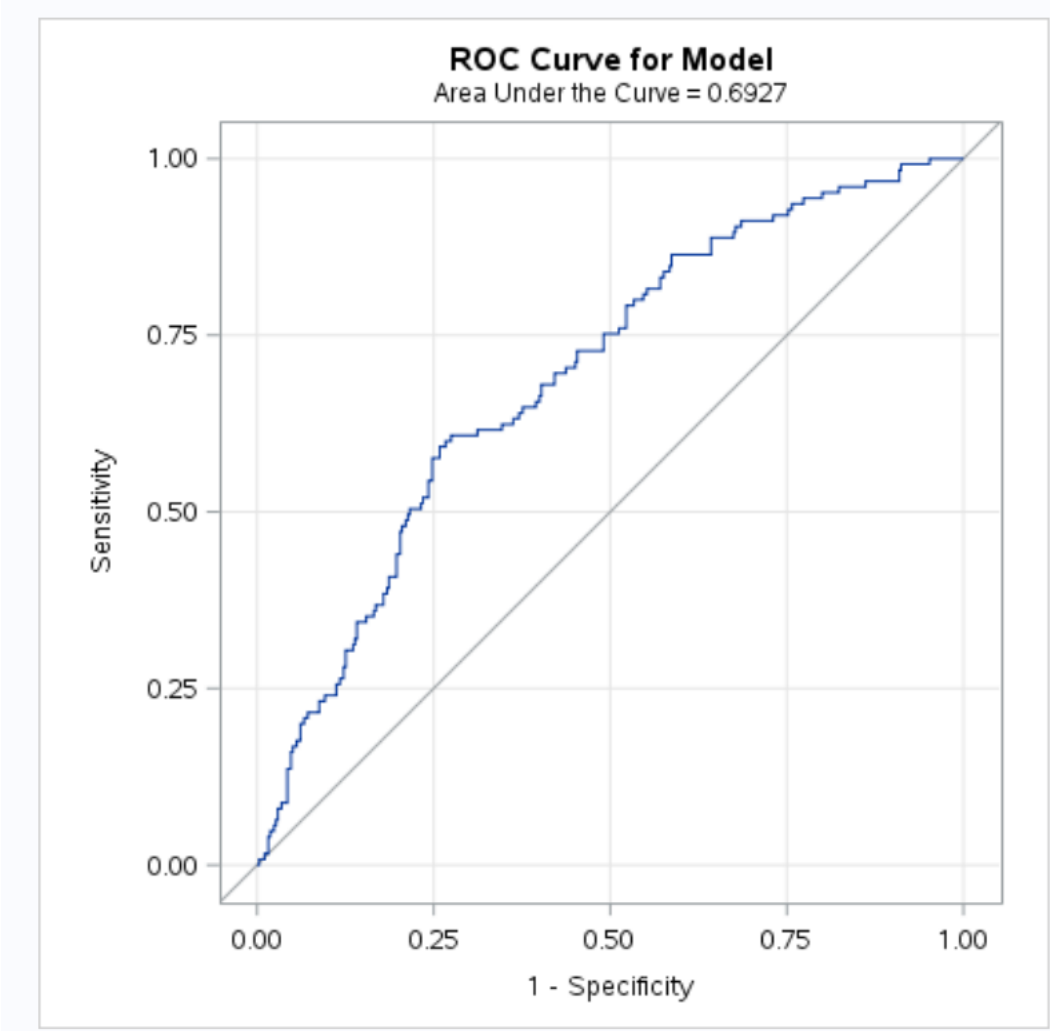
ROC Curve for Model
Area Under the Curve = 0.6927
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity

Figure 5:

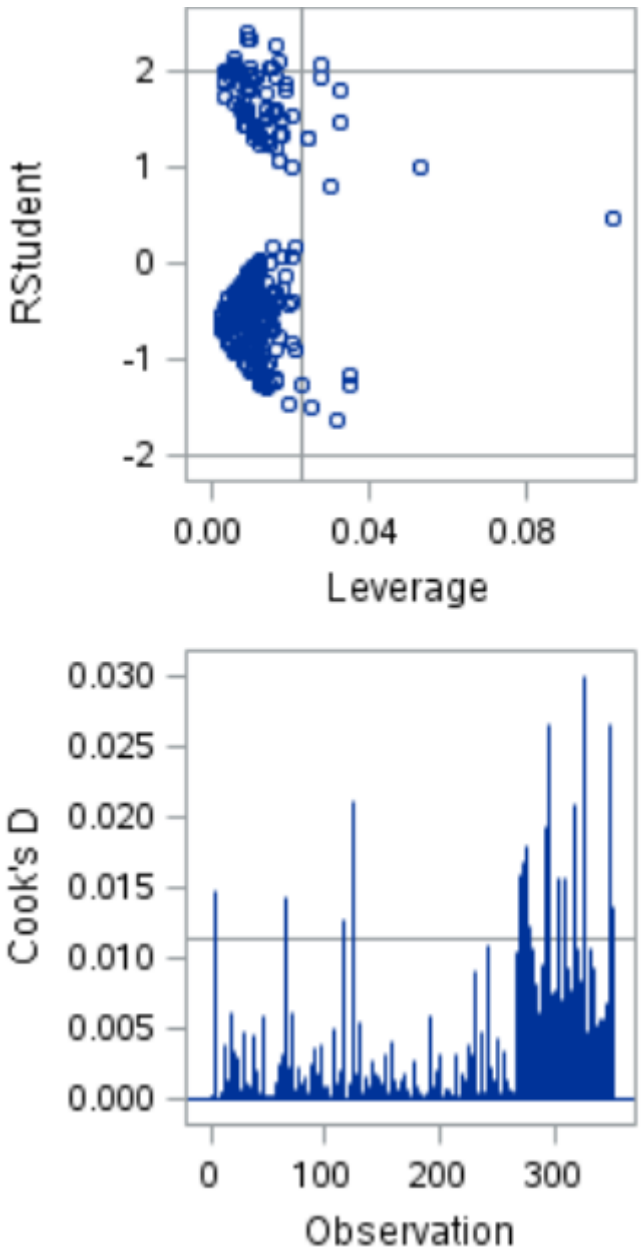

Figure 6:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7.25188	2.41729	14.54	<.0001
Error	347	57.67690	0.16622		
Lack of Fit	242	42.17690	0.17428	1.18	0.1660
Pure Error	105	15.50000	0.14762		
Corrected Total	350	64.92877			

Figure 7:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
HEIGHT	0.963	0.930	0.997
RATERISK	1.503	1.144	1.974
FRACSCORE	1.252	1.148	1.364

Figure 8:

Parameter Interpretation

The coeficient of each parameter in the model represents how much the natural logarithm of the odd ration of fracture changes if the parameter increases by one unit, assuming all other parameters remain unchanged. For example when RATERISK increases by 1, the log of odd ratio of frature increases by 0.4072, which means the od ratio increases by $e^0.4072 = 1.503 (this value along with it's 95% confidence interval is also present in the Odds ration Estimate table).

The negative coefficient of HEIGHT is consistent with negative correlation that was illustrated by a pink circle in the correlation matrix in the previous section.

Effect of each variable on the odds ratio (point estimate and 95% confidence interval) is presented in the figure 8:

Final conclusions from the analyses of Objective 1

In the above analysis we trained a logistic regression that predicts if a subject will have fracture in one year after the last examination. A stepwise parameter selection slected the following parameters as most predictive parameters: HEIGHT, RATERISK, FRACSCORE. The relative low deviance and high AUC indicates the predictive power of the model.

In addition, it appears that the model conforms to the assumptions of a logistic regression. There are no outliers that are too influential on the data and the lack of fit test lacked evidence to reject the fit of the model.

Addressing Objective 2

In this section we compare the logistic regression model against the following models: Linear Discriminat Analysis (LDA), Quadratic Discrimint Analysis (QDA), Random Forest (RF). To avoid overfiting and proerly evaluate generalization capability of the models we divided the data set into a train (60%) and test dataset (40%). We only used the three explanratory parameters that the stepwise parameter selection selected in the previous section (i.e. HEIGHT, RATERISK, FRACSCORE).

To compare the models, we present confusion matrix, acuracy, and AUC metrics. We also visually compare their ROC curves.

```
# confusion matrix

# logistic
table(test$FRACTURE, ifelse(test.predicted.glm < 0.5, 0, 1))
```

```
##
##       0   1
```

```
##   0 142   4
##   1  47   2
```

```
# lda
table(test$FRACTURE, test.predicted.lda$class)
```

```
##
##       0   1
##   0 140   6
##   1  41   8
```

```
# qda
table(test$FRACTURE, test.predicted.qda$class)
```

```
##
##       0   1
##   0 140   6
##   1  45   4
```

```
# random forset
table(test$FRACTURE,  test.predicted.rf)
```

```
##    test.predicted.rf
##       0   1
##   0 145   1
##   1  49   0
```

```
# accuracy rate

# logistic
mean(ifelse(test.predicted.glm > 0.5, 1 , 0) == test$FRACTURE)
```

```
## [1] 0.7384615
```

```
# lda
mean(test.predicted.lda$class == test$FRACTURE)
```

```
## [1] 0.7589744
```

```
# qda
mean(test.predicted.qda$class == test$FRACTURE)
```

```
## [1] 0.7384615
```

```
# random forest
mean(test.predicted.rf == test$FRACTURE)
```

```
## [1] 0.7435897
```

```
# Logistic regression AUC
prediction(test.predicted.glm, test$FRACTURE) %>%
  performance(measure = "auc") %>%
  .@y.values
```

```
## [[1]]
## [1] 0.7435001
```

```
# LDA AUC
prediction(test.predicted.lda$posterior[,2], test$FRACTURE) %>%
  performance(measure = "auc") %>%
  .@y.values
```

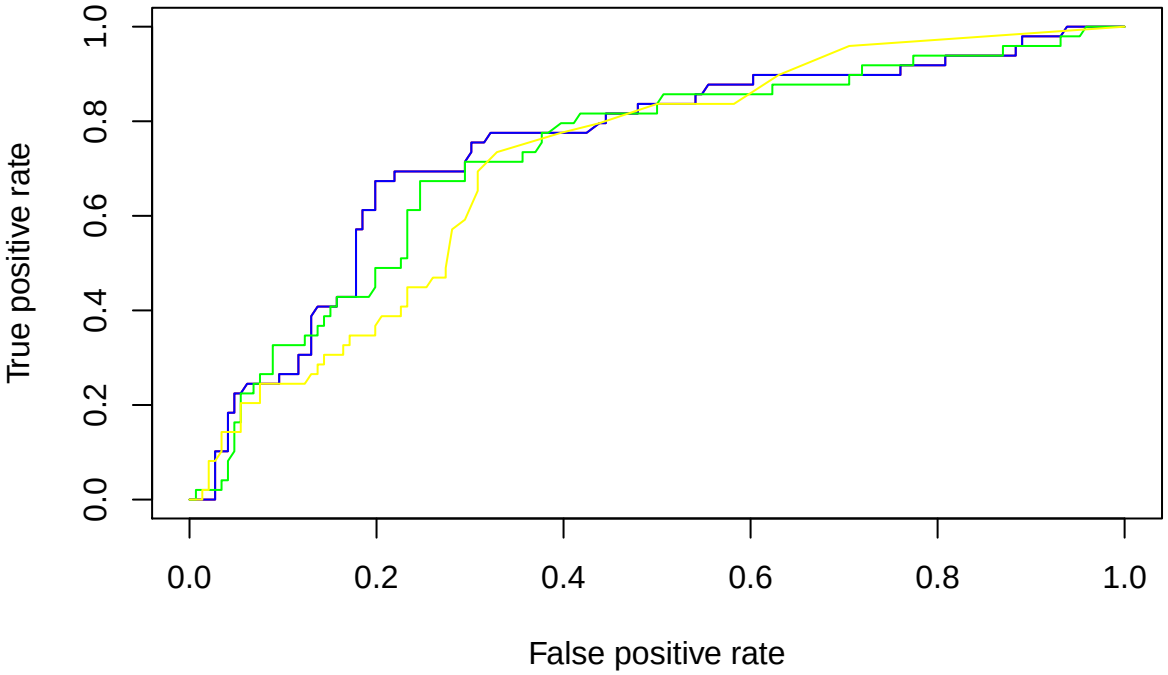

Figure 9: ROC curves

```
## [[1]]
## [1] 0.7435001
```

```
# QDA AUC
prediction(test.predicted.qda$posterior[,2], test$FRACTURE) %>%
  performance(measure = "auc") %>%
  .@y.values
```

```
## [[1]]
## [1] 0.7264467
```

```
# RandomForest AUC
prediction(test.predicted.rf.prob[,2], test$FRACTURE) %>%
  performance(measure = "auc") %>%
  .@y.values
```

```
## [[1]]
## [1] 0.7138664
```

Conclusion/Discussion Required

In this project we evaluated four different types of classification models for predicting probability of fracture using Glow500 dataset. First we use stepwise loggistic regression to find select parameters and then we trained four models using logistic regression, LDA, QDA, and random forest. The models were trained on only 60% of data and the rest of data was used for model evaluation. We looked at confusion matrix, accuracy rates, ROC curves and AUC to compare the models. Logistic regression and LDA have very similar performance characteristics and outpeformed QDA and random forest.

It appears that, using a logistic regression or LDA model, the likelihood of a woman having fractures can be predicted with some accuracy. However, even the most accurate models makes it apparent that the predictors used may be inadequate to predict with high certainty whether fractures will occur. However, since predictions

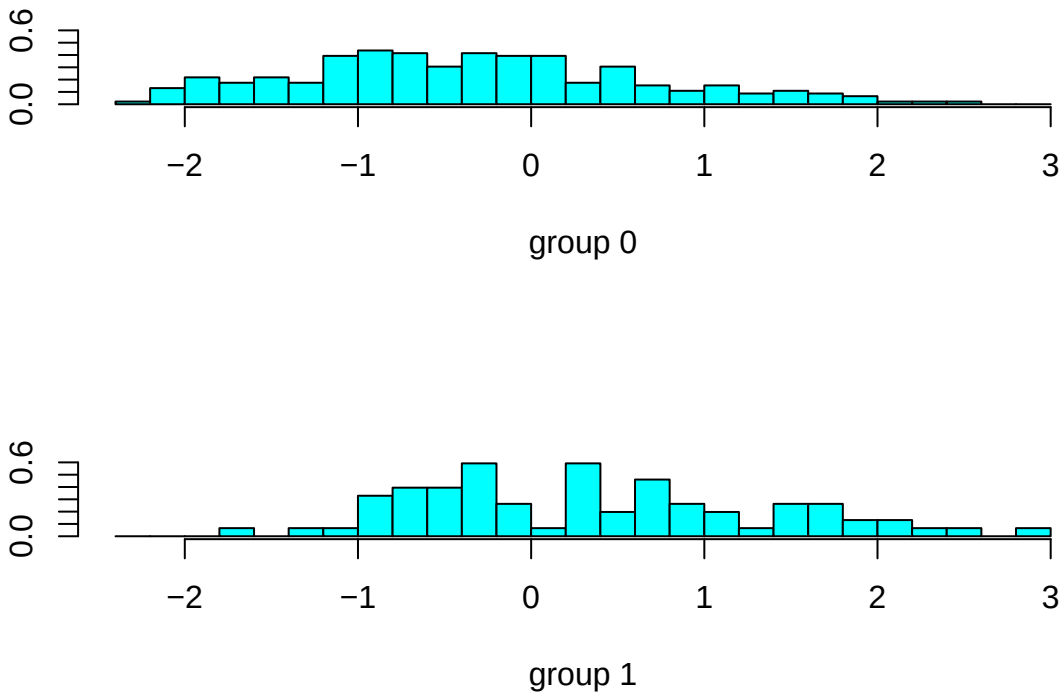

Figure 10: LDA model discriminative power

can be helpful to prevent injuries, a positive error could still be benefitial if an intervention/preventative treatment is inexpensive.

Appendix Required

R Codes:

```
# confusion matrix

# logistic
table(test$FRACTURE, ifelse(test.predicted.glm < 0.5, 0, 1))

# lda
table(test$FRACTURE, test.predicted.lda$class)

# qda
table(test$FRACTURE, test.predicted.qda$class)

# random forset
table(test$FRACTURE,  test.predicted.rf)
```

```
# accuracy rate

# logistic
mean(ifelse(test.predicted.glm > 0.5, 1 , 0) == test$FRACTURE)
# lda
mean(test.predicted.lda$class == test$FRACTURE)
# qda
mean(test.predicted.qda$class == test$FRACTURE)
```

```r
# random forest
mean(test.predicted.rf == test$FRACTURE)
```

```r
# Logistic regression AUC
prediction(test.predicted.glm, test$FRACTURE) %>%
  performance(measure = "auc") %>%
  .@y.values

# LDA AUC
prediction(test.predicted.lda$posterior[,2], test$FRACTURE) %>%
  performance(measure = "auc") %>%
  .@y.values

# QDA AUC
prediction(test.predicted.qda$posterior[,2], test$FRACTURE) %>%
  performance(measure = "auc") %>%
  .@y.values

# RandomForest AUC
prediction(test.predicted.rf.prob[,2], test$FRACTURE) %>%
  performance(measure = "auc") %>%
  .@y.values
```

SAS Codes:

```
LIBNAME MYSASLIB '/home/szarandioon0/';
DATA GLOW500_ORIG;
INFILE '/home/szarandioon0/statistics2/Project2/glow500.csv' DLM = ',' FIRSTOBS = 2;
INPUT SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK FRA
RUN;

DATA GLOW500(DROP = SUB_ID);
SET GLOW500_ORIG;
RUN;

proc factor data=GLOW500 simple corr;
run;

ods graphics on;
proc princomp data=GLOW500 plots(ncomp=3)=all n=5;
run;

title 'Stepwise Regression on Global Longitudinal Study of Osteoporosis in Women (GLOW) Dataset';
proc logistic data=GLOW500 outest=betas covout;
model FRACTURE(event='1')=SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT BMI PREMENO    MOMFRAC ARMASSIST
                    / selection=stepwise;
output out=pred p=phat lower=lcl upper=ucl predprob=(individual crossvalidate);
run;

proc logistic data=GLOW500 rocoptions(crossvalidate) plots(only)=roc;
model FRACTURE(event="1") = RATERISK FRACSCORE HEIGHT;
run;

data train test;
set GLOW500;
```

```
if rand('uniform') <= 0.3
then output test;
else output train;
run;

ods graphics on;
proc logistic data=train;
model FRACTURE(event="1") = RATERISK FRACSCORE HEIGHT / outroc=troc;
score data=test out=valpred outroc=vroc;
roc; roccontrast;
run;

proc logistic data=train plots(only)=roc;
model FRACTURE(event="1") = RATERISK FRACSCORE HEIGHT;
run;

proc logistic data=train rocoptions(crossvalidate) plots(only)=roc;
model FRACTURE(event="1") = RATERISK FRACSCORE HEIGHT;
run;

proc discrim data=train testdata=test canonical;
class FRACTURE;
var SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK FRACSCORE;
run;

proc hpforest data=train;
target FRACTURE/level=nominal;
input PRIORFRAC PREMENO MOMFRAC ARMASSIST SMOKE/level=nominal;
input SITE_ID PHY_ID AGE WEIGHT HEIGHT BMI RATERISK FRACSCORE/level=interval;
run;

/*
 * Checking assumptions
 */
proc reg data = train;
  model FRACTURE = FRACSCORE RATERISK HEIGHT;
  output out = t student=res cookd = cookd h = lev;
run;
```

References

- Wiley Series In Probability And Statistics, Section 1.6.3
- http://uc-r.github.io/discriminant_analysis