Statistical models and data analysis

Project 2

Abdirahman Mohamed Yassin

June 2019

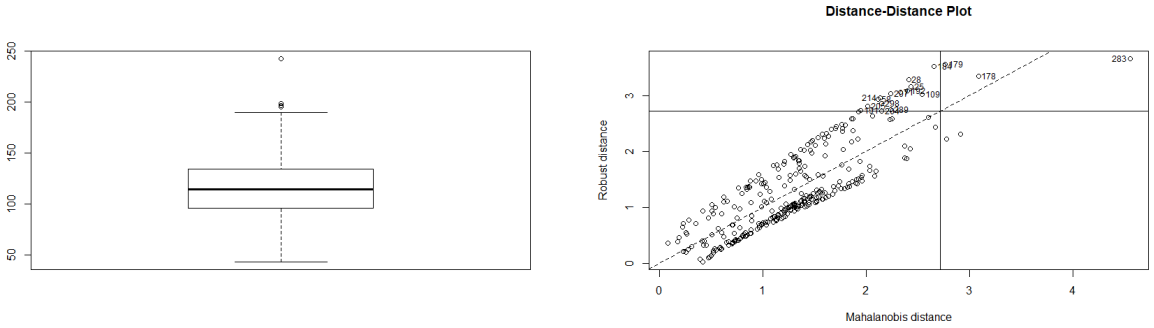
Contents

1 Exploratory data analysis

The data analysis is started by consulting the basic summary statistics of the diabetes dataset. Some obvious patterns can be deduced from the summary table. It is clear that the data are not centered. Additionally, some samples have wider variations than others while the observations of the non-categorical predictor variables BMI, BP and blood serum measurements have different units or metrics.

```
AGE              BMI              BP               S1              S2              S3              S4
 Min.   :19.00   Min.   :18.00   Min.   : 62.00   Min.   : 97.0   Min.   : 43.40   Min.   :22.00   Min.   :2.000
 1st Qu.:37.00   1st Qu.:23.07   1st Qu.: 85.00   1st Qu.:162.0   1st Qu.: 95.85   1st Qu.:40.00   1st Qu.:3.000
 Median :50.00   Median :25.60   Median : 93.00   Median :187.0   Median :114.50   Median :48.00   Median :4.000
 Mean   :47.94   Mean   :26.26   Mean   : 94.56   Mean   :188.1   Mean   :115.05   Mean   :49.49   Mean   :4.077
 3rd Qu.:59.00   3rd Qu.:29.00   3rd Qu.:105.00   3rd Qu.:211.0   3rd Qu.:134.30   3rd Qu.:57.00   3rd Qu.:5.000
 Max.   :79.00   Max.   :42.20   Max.   :131.00   Max.   :300.0   Max.   :242.40   Max.   :98.00   Max.   :9.090
       S5              S6               Y
 Min.   :3.497   Min.   : 58.00   Min.   : 31.0
 1st Qu.:4.259   1st Qu.: 83.00   1st Qu.: 84.0
 Median :4.630   Median : 92.00   Median :131.5
 Mean   :4.632   Mean   : 91.26   Mean   :147.3
 3rd Qu.:4.973   3rd Qu.: 98.00   3rd Qu.:200.2
 Max.   :6.105   Max.   :124.00   Max.   :341.0
```

Boxplots of the variables are plotted to further explore the skewness, variability of the observations and possible outliers. Most of the variables appear to be symmetric with few exceptions showing several observations falling behind the whiskers. A special case is observed in the boxplot of the the second blood serum measurement S2 with one observation lying outside the whisker at relatively higher distance and may indicate a possible outlier. Although not all outliers will have the same influence on the regression, robust analysis of S2 with the response variable showed the particular observation point is located furthest outside robust tolerance ellipse which corresponds to the outlying point from boxplot.



(a) **(b)**

Figure 1: Outlier detection, (a) boxplot, (b) mahalobonis vs robust distance

To obtain a good linear model, normality of the predictor variable and the response variable are evaluated and if necessary transformation is performed to approximate normal distribution. From the set of predictor variables that strongly deviate from normal distirbution (using Q-Q plot) BMI, S3 and S5 could be transformed transformed by the method of Box and Cox. Normal approximation after transformation was assessed using Shapiro Wilk test. The response variable showed strong deviation from normal distirbution. No appropriate transformation could be carried out and the analysis further proceeded without transformation of the response.

The scatter plot shows strong positive correlation between S1 and S2. Tabulated correlation values of the predictor variables further revealed a $\rho_{S1,S2}$ of 0.89 between the two predictor variables indicating high correlation. Consequently, multicollinearity can be an issue in this dataset and needs to be further examined. This is approached by the variance inflation factors (VIF) as the formal method to detect multicollinearity. The largest VIF value is 60 which is considerably larger than 10. Similarly, the mean of VIF is 14 which is also considerably larger than 1. Therefore, there is presence of multicollinearity and S2 is not considered for the regression analysis. Additionally, the scatter plot shows some correlation between the response and BMI, S5. This indicates a linear relationship between those predictors and the response. BMI and S5 are therefore expected to appear in the model.

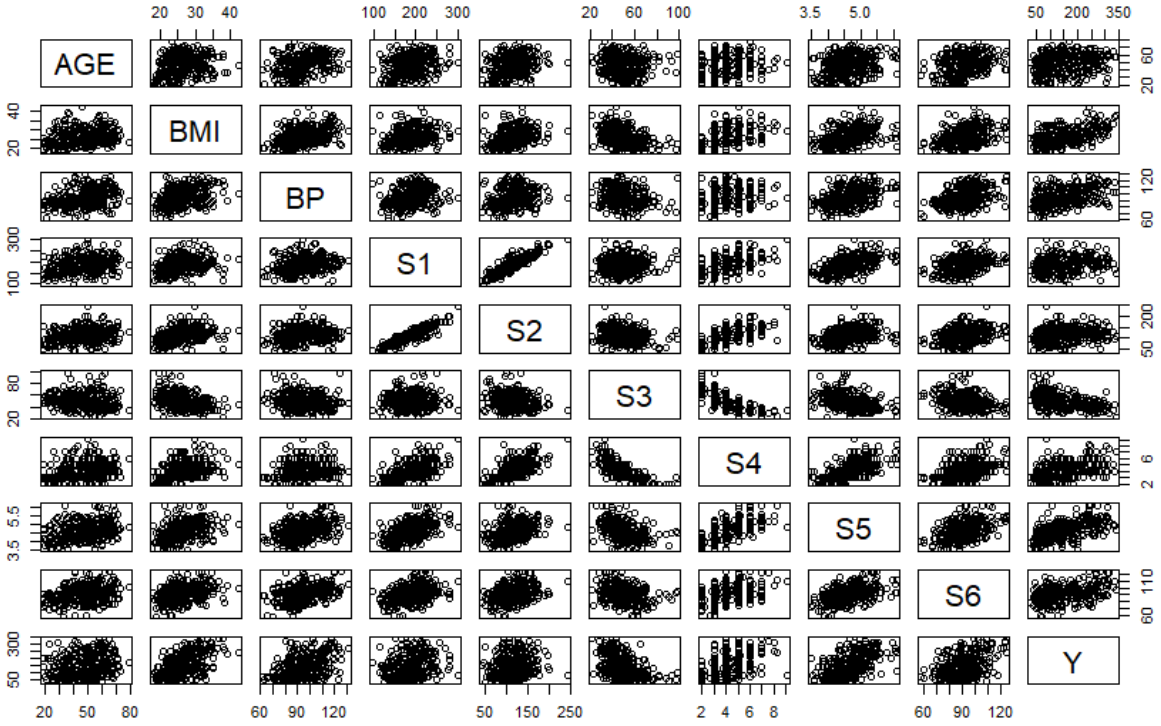


Figure 2: Scatter plot of predictors and response variable

2 Linear model

2.1 Variable selection

To construct the linear model for the problem, variable selection was performed using different methods. For this purpose, the data was divided into training set of 200 observations while the remaining is used for validation. The data consists of 9 non-categorical predictor variables and a categorical variable SEX coded as SEX = 1 for male and 0 for female.

Table 1: Variable selection

Criterion	Total variable selected
R_p^2	9
MSE_p	6
AIC	6

Both adjusted R_p^2 and AIC criterion resulted in the same variable selection leaving Age, S3 and S6 out of the model. Stepwise regression that combines both forward selection and backward elimination resulted in the same variable selection as the previous methods. Table shows the mean PRESS and the MSEP of the models based on the specified criteria.

Table 2: PRESS and MSEP for different criteria

Criterion	Mean CV-PRESS	MSEP
R_p^2	3056	2582
$\text{MSE}_p \& AIC$	2995	2560

From the table few remarks can be deduced. Based on the PRESS values, both models with 9 and 6 variables have similar performance with actual data (calibration) and the relatively similar MSEP indicates both models perform almost equally on the validation set. The MSE is 2750 and 2784 respectively which are not significantly different than MSEP. Therefore, both MSE and MSEP are equally reliable indicators for the predictive power of the both models. The full model resulted in adjusted R_p^2 of 0.53 and 0.52 for the reduced model.

2.2 Model assumptions

2.2.1 Normality

Figur 3 shows the normal quantile plot of the standardized residuals of the fitted model. No strong deviation from normality is observed. Shapiro Wilk test resulted in $p > 0.05$ on 5% significant level. As the standard residuals can be considered estimates for error terms, the normaility assumption of the error terms is therefore reasonable .

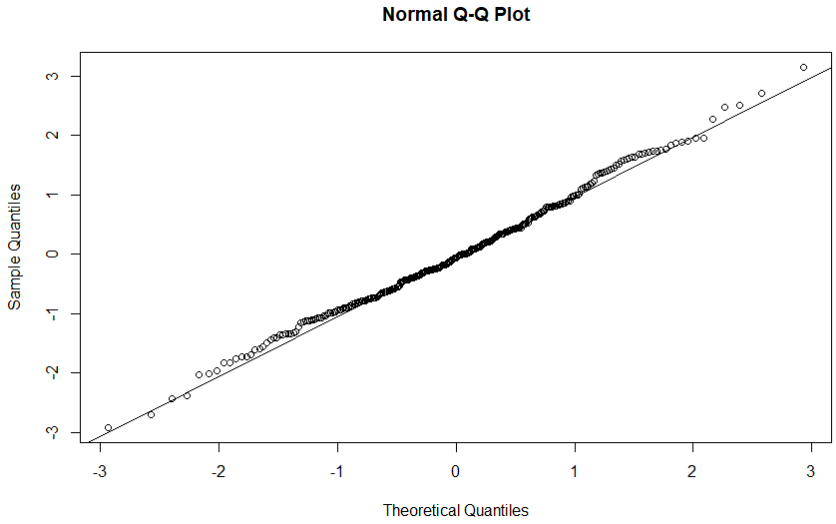


Figure 3: Q-Q plot of the standardized residuals

2.2.2 Residual plots

Figure 4 shows the residual versus index. It can be seen that there is no pattern in the horizontal band and therefore the error terms can be assumed to be uncorrelated. Furthermore four observations with residuals exceeding 2.5 can be identified.

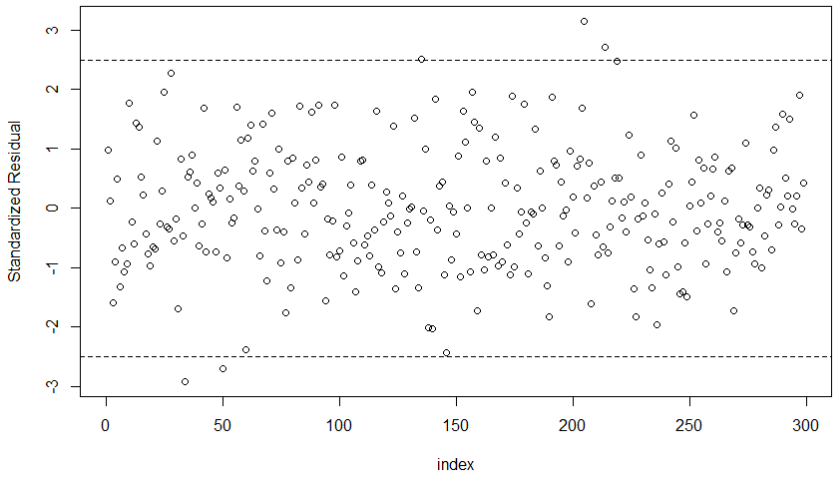


Figure 4: Standardized residual vs index

However , if we plot the residuals against the fitted values as shown in figure 5 some funnel like pattern can be observed where residuals are more spread out for some fitted values. This deviance from horizontal band may indicate violation of the homoscedasticity assumption of the model.

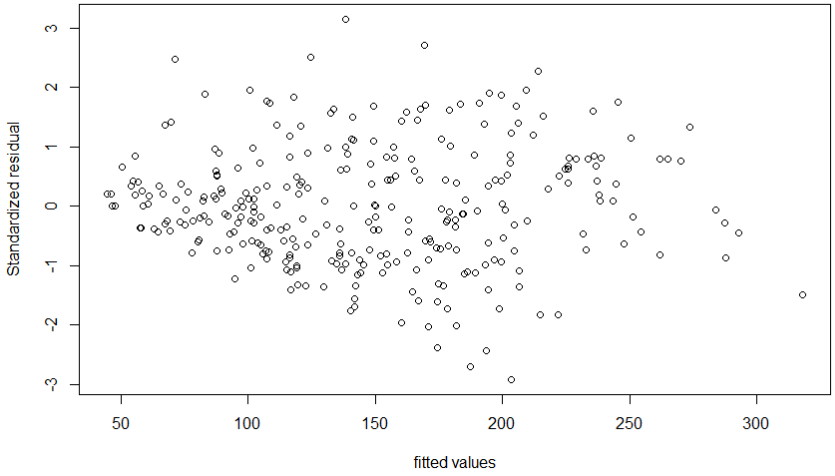


Figure 5: Fitted values vs Standardized residuals

Several phenomena may have led to the heteroscedasticity. One possible reason may be the nature of the response variable. Since it is quantitative measure for disease progression, it is possible that these are averages of some other measurement and therefore contribute with a non-constant variance. Transforming the response to more symmetric distribution did not alter the heteroscedasticity. Another aspect to look at how the residual varies with the individual predictors. Almost all the predictors, with the exception of S5 in the model showed similar funnel pattern demonstrating the heteroscedasticity. BMI, BP and S1 could be regressed to obtain estimates for the standard deviation and perform weighted regression using those estimes. However, the weighted regression didn't turn out to be a remedy for the heteroscedasticity as the same pattern was observed again.

2.2.3 Diagnostic plots

Figure 6 shows the diagnostic plot of the LTS regression (categorical variable excluded) as a tool to identify vertical outliers. Observation 135, 205, 214 and 209 were identified as vertical outliers that may affect the linear fit. However, LTS regression (categorical variable included) resulted in a similar model indicating that the data doesn't contain too many outliers that may have affected the model.

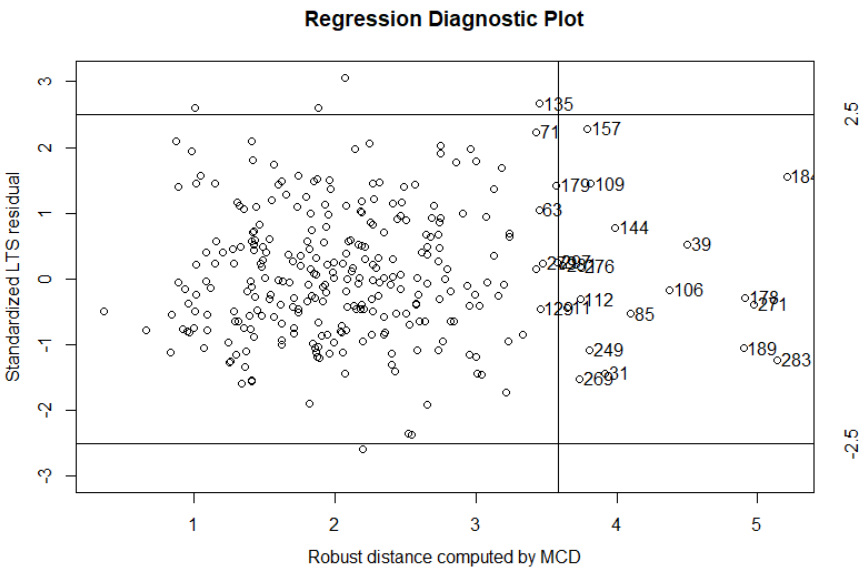


Figure 6: Diagnostic plot

3 Model insights

As mentioned above, transformation of the predictors and the response did not result in a model they better fits the assumption of homoscedasticity. As a consequence, the following model on which some insights are drawn consists of the original predictor variables and response variable making the interpretation more convenient. The obtained model is as follows

$$y = -310.17 - 26SEX + 5.16BMI + 1.25BP - 0.48S1 + 15.29S4 + 49.81S5$$

The model suggests male patients have better prognosis of less disease progression than female counter parts. From the first bloods serum measurement S1, it can be deduced that the disease may not not progress. On the other hand high BMI and BP are associated with increased disease progression. This association fits the general trend regarding diabetes in the medical sector. Similarly, measurements S4 and S5 have a positive influence on disease progression.