Report to investigate how well diabetes can be predicted in Pima Indian Women from 8 feature variables

Student 202078239

1 Introduction

The Pima Indians in Arizona have been studied since 1965 for having the highest rates of type 2 diabetes in the world. Type 2 diabetes is a condition that causes glucose levels in the blood to become too high and can lead to life-threatening complications. According to the Centres for Disease Control and Prevention, diabetes is the 7th most common cause of death in the United States. Obesity was reported in Pima women as early as 1775,[1] suggesting there's an important genetic factor, but the prevalence of obesity and diabetes has increased rapidly since the Pima were exposed to Western culture including a high fat and sugar diet and a more sedentary lifestyle. Increasing body weight and increasing exposure to diabetes 'in utero' has started a vicious cycle where it is now common for Pima children as young as 10 to be affected.[2] It is crucial that we find the best way to tackle this diebetes pandemic in order to improve the health of future generations, for both Pima Indians and people worldwide.

1.1 Data

In this report we explore data on female Pima Indians over 21 years old, adapted from the PimaIndiansDiabetes2 dataset in the mlbench package, and investigate which factors lead to an increased risk of having type 2 diabetes. Data on the following variables are known and investigated:

pregnant Number of times pregnant

glucose Plasma glucose concentration

pressure Diastolic blood pressure in mm Hg

triceps Triceps skin fold thickness in mm

insulin 2-Hour serum insulin in μU/ml

mass Body mass index, BMI, measured as weight in kg/height in m^2

pedigree Diabetes pedigree function, DPF, a score of family history of diabetes

age Age in years

diabetes Test for diabetes – positive or negative

There are 8 further variables in the data which are scaled versions of the feature variables, with names ending in .sc. These variables have been transformed by subtracting its mean and dividing by its standard deviation. This minimises the effect of any outliers when a distance based algorithm is used and gets close to a standard normal distribution with mean 0. There are observations from 391 women in the data frame and all variables except diabetes (response variable) are numeric. Diabetes is a binary factor depicting whether they test positive or negative for diabetes. There are no NAs or missing data, and all variables have appropriate ranges of data. Equally, there are no obvious recording errors or codes used to replace missing data, so the data are considered clean and ready to use.

2 Method

Firstly, summary statistics and appropriate plots (barcharts, histograms and boxplots) were explored to understand the data. Potential issues of scale and high collinearity between the feature variables were investigated through covariance and correlation matrices. Transformations of skewed variables were made to create more symmetric data to help with model fitting. The dataset was then randomly split up into a training and test set with a ratio 3:1, and logistic regression modelling with appropriate variable selection was carried out.

Logistic regression is a generalised linear model used to link a binomial response variable to a set of feature variables. We want to predict the probability of an outcome (diabetes), which is measured from 0 to 1. However, the linear part of the equation has values between $-\infty$ to ∞ so we must ensure that both sides of the equation are balanced. If we consider the odds ratio rather than probability we have a scale of 0 to infinity. If instead we take the (natural) log odds of the outcome, we have a scale of $-\infty$ to ∞, which balances the equation. We therefore predict the log odds of our binary outcome (diabetes) using a linear equation of the feature variables and intercept, equation 1, where π is the probability of "success" and logit is the link function.

$$\text{logit}[\pi(x)] = \ln[\pi(x)]/[1-\pi(x)] = \beta_0 + \beta_1(x) \tag{1}$$

The parameters are estimated by maximum likelihood estimation using iterative numerical methods so it requires a large sample size. Assumptions made include observations being independent; little or no multicollinearity among the independent variables; and a linear relationship between the transformed response and the feature variables, shown in Appendix A. All assumptions were considered to be valid.

The fit of the final model was then assessed using Akaike's Information Criterion, AIC and the residual deviance values, found in the model summary, both of which we want to be as small as possible. $AIC = -2(\text{log-likelihood}) + 2k$, where k is the number of parameters in the model, so AIC favours the smaller model when the log-likelihoods are similar. The log-likelihood is the log-likelihood ratio of a model to a fully saturated model, with as many parameters as observations (full model). The test statistic for comparing a simpler model with a more complex one is the difference in their deviances and this follows a chi-square distribution and requires large samples. The anova function was used to get a chi-square test to statistically check whether the smaller model fitted sufficiently well compared to the full model. A receiver operating characteristic (ROC) curve of the model, which plots sensitivity against 1-specificity, was then used to find a suitable cut-off value for the final classification model and the performance of the classifier was calculated using sensitivity, specificity and correct classification values.

Finally, principal components analysis, PCA, was performed on the original, unscaled data and compared with the logistic regression model. PCA is useful when there are lots of numerical variables and when the variables are correlated. It combines variables into a smaller number of orthogonal principal components using loadings, which give a weighting to each variable in the component, thus simplifying interpretation. The scores are the linear combinations of the data that are determined by the coefficients for each principal component.

The expected value of variable j for observation i is shown in equation 2.

$$E[x_{ij}] = \phi_{j1}\nu_{i1} + \phi_{j2}\nu_{i2} + ... + \phi_{jk}\nu_{ik} \tag{2}$$

where ϕ_{jh} are the loadings of variable j on principal component h and ν_{ih} is the score of principal component h for observation i.

PCA is carried out using eigen-decomposition of either the covariance or the correlation matrix of the variables. If all the variables are on the same scale the covariance should be used, otherwise the correlation matrix is preferred. Since the data are not all on the same scale, the PCA was calculated from the correlation matrix, using the prcomp function in R. The loadings were viewed in the PCA summary and scores were calculated using the predict function.

Results from the logistic regression and the principal components analysis were compared to check for agreement.

3 Results

3.1 Exploring the data

In figure 1 we see that about 1/3 of Pima women in the study have tested positive for diabetes (130/391). Looking at the summary data in table 1 and the histograms and bar chart in figure 2, we see that the number of pregnancies ranges from 0 to 17 with a median of 2 and it is most common for the women to have had 1 or 2 pregnancies. The bar chart is highly skewed by some very high numbers of pregnancies. Glucose concentration ranges from 56.0 to 198.0 units with a median of 119.0 and the most common glucose range is 80 to 100 units. Blood pressure ranges from 24.0 to 100.0 mm Hg with a median and mean of about 70 mm Hg. These data are vaguely symmetric with long tails. Triceps skin thickness ranges from 7.0 to 63.0 mm with a median and mean of about 29 mm. These data are more symmetric than the other variables. Serum insulin ranges from 14.0 to 846.0, covering a huge range of values. It's median is 125.0 μ U/ml with very skewed data. Its interquartile range is 76.5 to 190.0 so there are some extreme values found in this variable. BMI values range from 18.2 to 67.1 kg/m^2 with a median and mean of about 33 kg/m^2. For comparison, a BMI of about 18 to 25 kg/m^2 is considered to be a healthy range in the UK. These data are also skewed due to some very high values above 50 kg/m^2. The diabetes pedigree function data is also skewed to the right due to some very high values. The median is 0.447 in a range of 0.085 to 2.42 units. Finally, the age of the women ranges from 21 to 81 years with a median of 27. Again data is highly skewed to the right with the most common age being 21 to 25 years old.

Table 1: Data Summary

pregnant	glucose	pressure	triceps	insulin
Min. : 0.000	Min. : 56.0	Min. : 24.00	Min. : 7.00	Min. : 14.0
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 62.00	1st Qu.:21.00	1st Qu.: 76.5
Median : 2.000	Median :119.0	Median : 70.00	Median :29.00	Median :125.0
Mean : 3.302	Mean :122.6	Mean : 70.62	Mean :29.12	Mean :155.9
3rd Qu.: 5.000	3rd Qu.:143.0	3rd Qu.: 78.00	3rd Qu.:36.50	3rd Qu.:190.0
Max. :17.000	Max. :198.0	Max. :110.00	Max. :63.00	Max. :846.0

mass	pedigree	age	diabetes
Min. :18.20	Min. :0.0850	Min. :21.00	neg:261
1st Qu.:28.40	1st Qu.:0.2695	1st Qu.:23.00	pos:130
Median :33.20	Median :0.4470	Median :27.00	
Mean :33.07	Mean :0.5226	Mean :30.87	
3rd Qu.:37.05	3rd Qu.:0.6845	3rd Qu.:36.00	
Max. :67.10	Max. :2.4200	Max. :81.00	

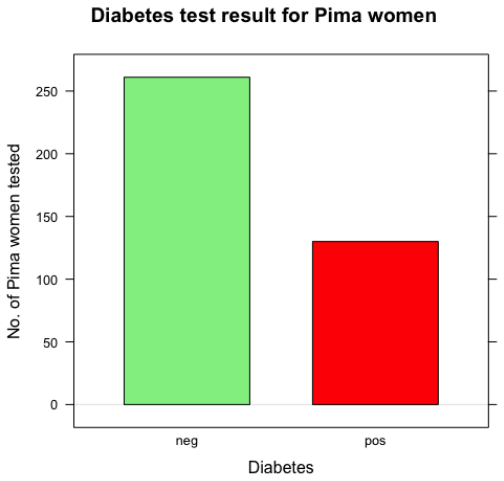

Figure 1: Barplot showing diabetes test results of the women

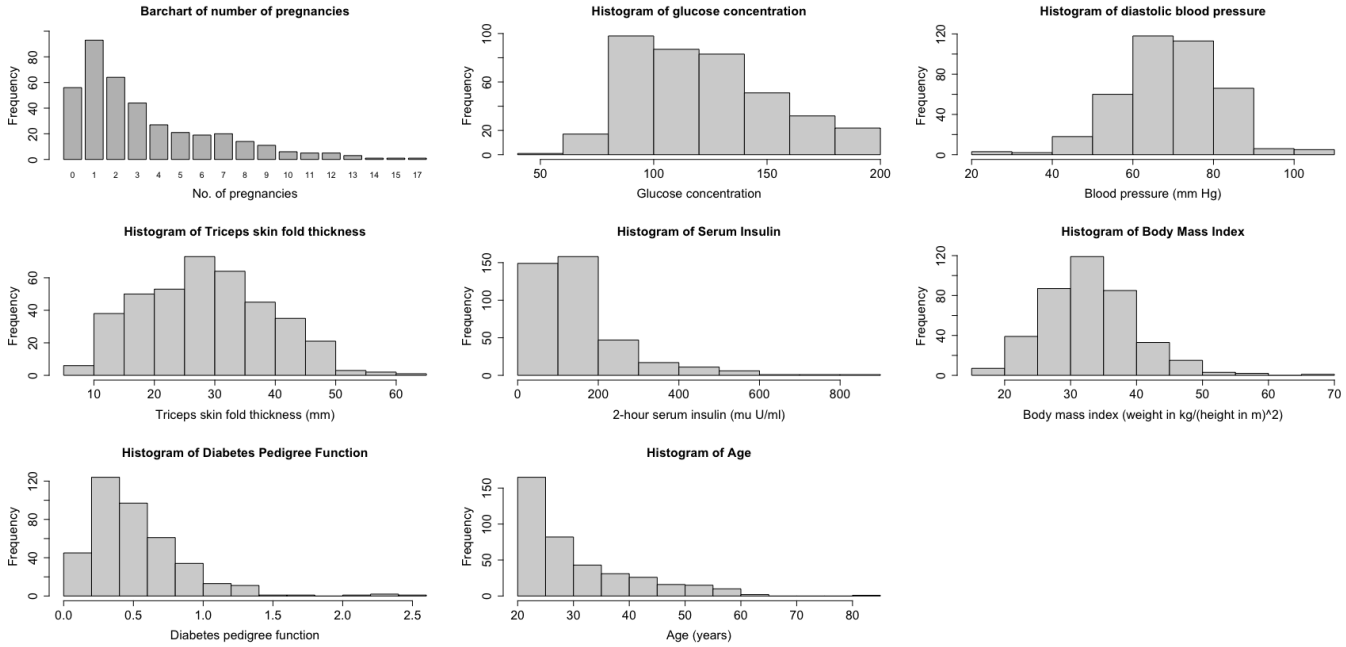

Figure 2: Summary graphs of the data collected for each variable

Figure 3 shows the observations of each feature variable for women who tested positive for diabetes compared to those who tested negative. Generally, those with diabetes had higher readings for all the feature variables. Every plot shows some outliers, shown as circles, but plots of the insulin, BMI, DPF and age all have a large number of outliers which would disproportionally affect our model, so scaled data were used to reduce the effect of these.

Continuing to look at figure 3 it can be seen that the spread of the pregnancy data is larger for those with diabetes and skewed towards higher numbers of pregnancies, although there are some outliers in the negative group with a high number of pregnancies. Almost the entire interquartile range for glucose concentration is higher in the positive group for diabetes suggesting that this variable will be very significant in classifying who is likely to have diabetes. The spread of blood pressure data is similar for both those with and without diabetes, but the range of data for those with diabetes are shifted towards higher blood pressure readings. Also, triceps skin fold thickness looks like being a good indicator of diabetes as the 1st quartile level of data for those with diabetes is almost aligned with the median value for non-diabetics. For levels of serum insulin the spread of data is slightly larger and the median is higher in diabetics. There are many outliers for the serum insulin levels in both groups. BMI for diabetics has a higher median and shows a smaller spread but with more outliers than non-diabetics. The DPF again shows diabetics with a higher median and the spread of data is slightly larger also. Finally, the interquartile range of age for diabetics is entirely above the median of the non-diabetic group, suggesting that younger women are less likely to have diabetes. However there are many outliers denoting those who are older but in the non-diabetic group.

Table 2 shows the covariance matrix of the feature variables. The absolute values of covariance range from about 0.6 to just over 14000. The difference in scales of the variables is huge so in order to compare any strength of association, we must standardise the data. Correlation standardises the covariance resulting in a scale of -1 to 1, figure 4. For logistic regression model fitting, it would be sensible to use the scaled data to avoid the large magnitude variables over-shadowing others.

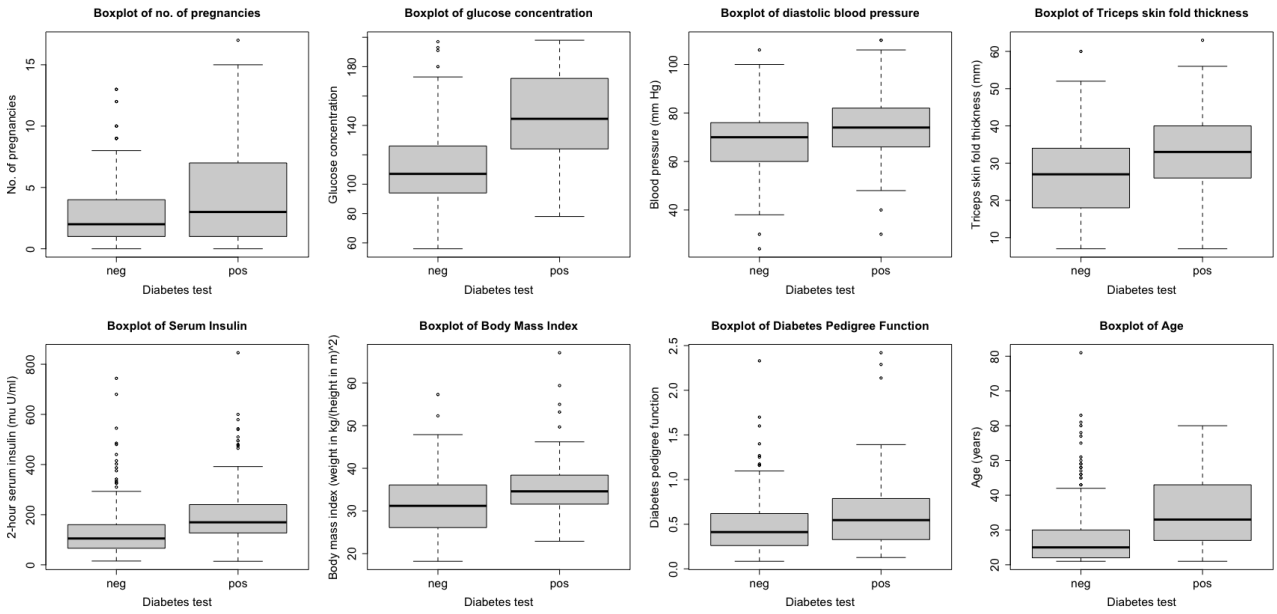

Figure 3: Boxplots showing relationships between diabetes test and each of the feature variables

Table 2: Covariance matrix

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
pregnant	10.339	20	8.60	3.2	30	-0.6	0.009	22.3
glucose	19.705	955	81.05	64.6	2136	45.5	1.497	108.5
pressure	8.597	81	155.78	30.1	143	26.5	-0.077	38.5
triceps	3.165	65	30.11	110.5	226	49.0	0.579	18.2
insulin	30.283	2136	143.15	225.9	14144	188.3	5.558	264.6
mass	-0.569	46	26.52	49.0	188	49.4	0.384	5.1
pedigree	0.009	1	-0.08	0.6	6	0.4	0.120	0.3
age	22.317	108	38.52	18.2	265	5.1	0.302	104.3

Figure 4 shows that there are 3 moderately strong correlations among the feature variables. These are glucose and insulin levels (0.58); BMI and triceps skin fold thickness (0.66); and age and number of pregnancies (0.68). None are above 0.7 so multi-collinearity shouldn't be a major issue, but the values are high enough to give some concern. I suspect we may not want both age and pregnancy in a model as women with many pregnancies are likely to be older anyway. Similarly we expect those with a high BMI to have higher triceps skin fold thickness also, so again, we may choose to only include one of these in a model.

3.2 Logistic Regression

Before constructing a regression model, the data were randomly split into a training and test set using the rbinom function, producing 298 observations in the training set and 93 in the test set. A logistic regression model was produced using the training data and all the scaled feature variables to predict the log likelihood of having diabetes. A further model was produced using the transformed, scaled variables to see if these improved the model. The residual deviance for the untransformed model was 273.58 compared to 267.39 for the transformed model. These values are not very different so the untransformed data were used to avoid unnecessary complexity. There were variables considered to be insignificant, with high p-values so a backwards selection process was used, removing variables with the highest p-value, one at a time, to reduce the model to include only variables with p-values that are all significant to 0.05. First triceps.sc was removed, then pressure.sc, insulin.sc and age.sc. This produced a final model involving the scaled pregnancy, glucose, mass and pedigree variables, shown in equation 3. The coefficients and p-values are given in table 3.

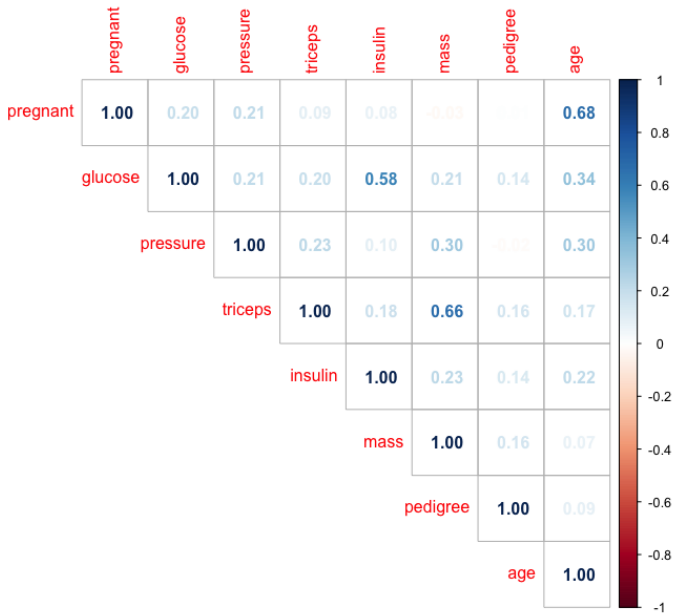

Figure 4: Correlation matrix

$$\begin{aligned}\text{Log odds ratio (diabetes)} = & -0.9642 + 0.5441 \text{ pregnant.sc} + 0.9895 \text{ glucose.sc} \\ & + 0.6617 \text{ mass.sc} + 0.4247 \text{ pedigree.sc}\end{aligned} \tag{3}$$

Table 3: Coefficients and associated p-values

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9642     0.1590  -6.064 1.33e-09 ***
pregnant.sc   0.5441     0.1488   3.657 0.000255 ***
glucose.sc    0.9895     0.1596   6.199 5.69e-10 ***
mass.sc       0.6617     0.1724   3.837 0.000124 ***
pedigree.sc   0.4247     0.1687   2.518 0.011813 *
```

The lowest p-value for these coefficients is for glucose suggesting that this is the most important variable and it has the highest association with having diabetes. On the left hand side of the equation we have the log odds ratio of having diabetes and on the right hand side we have the intercept and slope parameters of the feature variables. The intercept of -0.9642 is the log odds of having diabetes when all the feature variables are zero. The slope coefficients of the feature variables give the change in log odds associated with a unit change in that variable. For example for every additional pregnancy a Pima woman has, the log odds of having diabetes increases by an additive factor of approximately 0.5. If any of these slope parameters were negative then they would reduce the log odds of having diabetes. However, all the variables in our model are positive and therefore increase the log odds. The 95% confidence intervals of the coefficients for the log odds ratio are given in table 4.

Table 4: 95% Confidence intervals for log odds ratio

		2.5%	97.5%
(Intercept)	-0.9641703	-1.27582320	-0.6525174
pregnant.sc	0.5440969	0.25246722	0.8357265
glucose.sc	0.9895059	0.67664363	1.3023682
mass.sc	0.6617385	0.32375271	0.9997243
pedigree.sc	0.4246962	0.09407903	0.7553133

The confidence intervals for the coefficients of the log odds are all approximately +/-0.3. None span zero confirming that all our model parameters are significant. We can take the exponent of these values to get the confidence intervals for the odds ratio also, table 5. The odds of having diabetes increase by a multiplicative factor of the estimates of the variables. So for a unit increase in glucose levels, the odds of having diabetes increase by about 2.7 times.

Table 5: 95% Confidence intervals for odds ratio

		2.5%	97.5%
(Intercept)	0.3812994	0.279201	0.5207332
pregnant.sc	1.7230515	1.287197	2.3064892
glucose.sc	2.6899050	1.967264	3.6779964
mass.sc	1.9381589	1.382305	2.7175326
pedigree.sc	1.5291257	1.098647	2.1282782

The residual deviance for the final model is 276.62, which is similar to the full untransformed model value of 273.58, shown in table 6. The AIC, given in the model summary (not shown), is 286.62 for the final model, lower than the AIC of 291.58 for the full model. This is expected when the log-likelihood is similar for both models as AIC favours a smaller model, as mentioned previously. The p-value for the chi-square test, comparing the final and saturated model is 0.552, which is >0.05 significance level. Therefore we conclude that there is no significant difference in deviance between the models, and that the final (smaller) model can be considered a good fit.

Table 6: Comparing deviances of final model and saturated model

```
Analysis of Deviance Table

Model 1: diabetes == "pos" ~ pregnant.sc + glucose.sc + mass.sc + pedigree.sc
Model 2: diabetes == "pos" ~ pregnant.sc + glucose.sc + pressure.sc +
    triceps.sc + insulin.sc + mass.sc + pedigree.sc + age.sc

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       293     276.62
2       289     273.58  4   3.0352    0.552
```

Figure 5 shows the ROC curve using the training data, to choose the best cut-off value for the classifier. The best cut-off given by ROC is -0.881. Using -0.881 as the cut-off, we get a classification matrix for the training data, table 7.

The correct classification rate is $(234/298) \times 100 = 78.5\%$. This is the percentage who were correctly classified by the training model. Sensitivity is $(84/102) \times 100 = 82.4\%$. This is the percentage of those with diabetes that were correctly classified to have diabetes. Specificity is $(150/196) \times 100 = 76.5\%$. This is the percentage of those without diabetes that were correctly identified to not have diabetes.

Ideally, we would like these values to be above 80% but they are fairly close to that. Changing the cut-off value used will affect the sensitivity and specificity. However, if we increase sensitivity,

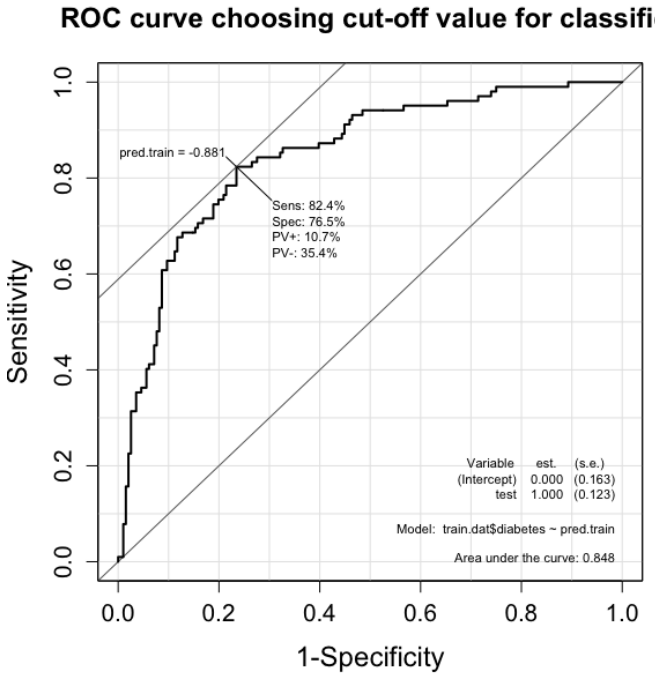

Figure 5: ROC curve to choose a cut-off according to required balance of sensitivity and specificity

specificity will decrease. It is useful to find a good balance appropriate to the study. A model with higher sensitivity will miss fewer cases that are positive, but if the treatment would be harmful to someone who is negative, then the specificity needs to be high also. For mild diabetes a key treatment is modifying someone's lifestyle by increasing exercise and improving their diet, neither of which would be harmful to someone who was negative. So for this study it would be reasonable to choose a cut-off value that leads to a higher sensitivity than specificity. However, looking at figure 5, if we increase the sensitivity further from our cut-off, the specificity decreases comparatively more, so it would not really be sensible to do that in this case.

Table 7: Classification matrix for training data

	neg	pos
FALSE	150	18
TRUE	46	84

We then tested our model on the unseen test data with the same cut-off chosen using the training data, to see how good the model is at predicting the likelihood of having diabetes in a new dataset. Table 8 shows the classification matrix for the unseen test data. The correct classification is slightly better for the test data at 79.6%. The sensitivity is marginally worse at 82.1% and the specificity is slightly better at 78.5%. These values suggest that the model is doing a good job of classifying who is likely to get diabetes.

Table 8: Classification matrix for test data

	neg	pos
FALSE	51	5
TRUE	14	23

3.3 Principal component analysis

Next, principal component analysis, PCA, was used and compared with the logistic regression model. The summary of the PCA is shown in table 9.

Table 9: Summary from principal component analysis function

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8
Standard deviation     1.5995 1.2469 1.0955 0.9781 0.84931 0.63342 0.55812 0.54415
Proportion of Variance 0.3198 0.1943 0.1500 0.1196 0.09017 0.05015 0.03894 0.03701
Cumulative Proportion  0.3198 0.5141 0.6642 0.7837 0.87390 0.92405 0.96299 1.00000
```

We can see that about 32% of the variation is explained by PC1 and a further 19% by PC2. Ideally we want to explain 75 to 80% of the variability, so 4 principal components would be needed from inspecting the cumulative proportions in table 9. We can visually check this with a scree plot, figure 6, but there is no obvious elbow in the plot to indicate a good choice for the number of components to use. Instead, we could use Kaiser criterion which requires the use of correlations in the PCA. Kaiser criterion uses all components with a variance above 1. This would lead us to use 3 principal components. However, since this would only account for about 66% of the variation, using 4 components would be better here.

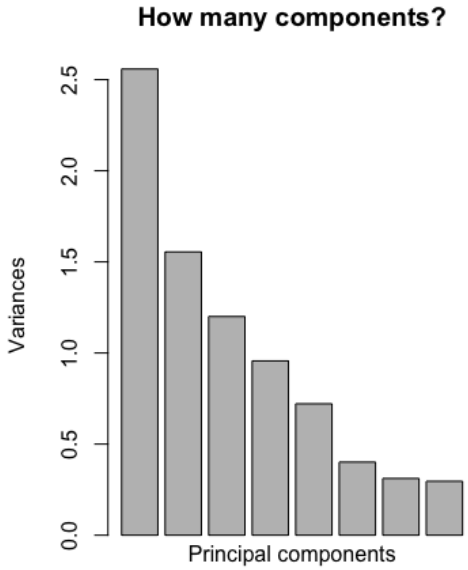

Figure 6: Scree plot to choose how many principal components to use

Table 10 shows the loadings for the first 4 principal components only. These loadings are proportional to the correlations of the variables with the given principal component. Any loadings <0.2 can therefore be ignored. PC1 shows an average as all values are positive. Pedigree is <0.2 and all other variables have fairly similar loadings of between about 0.32 and 0.42. PC2 has positive and negative loadings, showing a contrast. Glucose, blood pressure, insulin and pedigree are all <0.2. The contrast is therefore between numbers of pregnancies and age against triceps skin fold and BMI. PC3 again is a contrast. This time it is the number of pregnancies, blood pressure, triceps skin fold and BMI against glucose, insulin and pedigree. Insulin is the most (negatively) associated variable with PC3 at -0.58. PC4 is another contrast. Pregnancies, triceps skin fold, BMI and age are all <0.2, so the main contrast

of PC4 is between glucose, blood pressure and insulin against pedigree. Pedigree is highly (negatively) associated with PC4, with a value of -0.845.

Table 10: Loadings for the first four principal components

```
Rotation (n x k) = (8 x 8):
               PC1         PC2        PC3         PC4
pregnant 0.3159315 -0.55094188  0.2165431 -0.19838027
glucose  0.4246854 -0.04435920 -0.4730645  0.22793518
pressure 0.3288821 -0.03475099  0.3909728  0.30780668
triceps  0.3818751  0.41432566  0.3031794 -0.09074290
insulin  0.3643549  0.06902972 -0.5820665  0.25543167
mass     0.3666740  0.50618932  0.2576116  0.03266386
pedigree 0.1664269  0.18890269 -0.2562666 -0.84536209
age      0.4134433 -0.47428363  0.1170636 -0.15736031
```

The biplot in figure 7 shows the scores of PC1 (Dim1) and PC2 (Dim2) plotted with the loadings (red lines) of the variables on the principal components. The magnitude of the loadings is represented by the length of the arrows. Variables with small, acute angles with the x-axis are highly associated with PC1. These include blood pressure, glucose and insulin, which are positively related with PC1. All variables show positive association with PC1. Variables with the smallest angles to the y-axis are those that are most highly associated with PC2. BMI and triceps are the most positively associated variables with PC2 and age and number of pregnancies are the most negatively associated with PC2, which agrees with our interpretation from table 10. Blood pressure, glucose and insulin are close to being orthogonal to PC2 meaning that they have little contribution to PC2.

It can also be seen that some variables are highly correlated. Blood pressure and glucose are on the same trajectory so are very highly correlated. Similarly pedigree is almost the same as triceps skin fold. It can also be seen that BMI, triceps skin fold and pedigree have very little association with the number of pregnancies and age. There are a few outlying observations shown around the edges of the plot. Observation 226 is one and can be seen to have high values of both PC1 and PC2. Observation 111 has a high value of PC1 and a moderately high value of PC2. Observation 14 however, has a large, negative value of PC2 and only a moderately positive value of PC1. Finally, it should be noted that observations that a variable is pointing towards have high values of that variable. For example, 120, 207 and 388 have high values for triceps skin fold.

In the biplot, there are 3 general directions that the loadings are pointing in; along the positive PC1, towards positive PC2 and towards negative PC2. The logistic regression model includes the variables of pregnancies, glucose, BMI and pedigree. There is a variable in the model from each of the 3 general directions of the loadings. BMI and pedigree are on similar trajectories on the biplot, but the other 3 variables are all as spread out as they could be with the largest angles between them as possible. Interestingly, pedigree was the last variable to be kept in the model selection and it has the lowest weighting so I wonder if removing this variable from the regression model would not lose too much information. We can also compare the variables in the model to the loadings in table 10. Here we see that the largest weighting in PC4 comes from pedigree. In PC3 the largest weighting comes from insulin and glucose. These variables are highly correlated as shown in figure 4, so it is reasonable to only include one in the model. In PC2 the highest weightings come from pregnancy and mass. Finally, the highest weighting in PC1 comes from glucose. So, the variables used in the regression model are all found to be those with the highest loadings in the principal components.

Figure 7: Biplot showing scores and loadings for PC1 and PC2

4 Conclusion and Discussion

Analysis of the dataset using logistic regression and principal component analysis appears to provide us with comparable results that are in agreement with each other. Our analysis provides evidence to suggest that the most relevant feature variables in predicting the outcome of someone testing positive or negative for diabetes are plasma glucose concentration levels, BMI score, number of pregnancies and diabetes pedigree function score.

Diabetes pedigree function plays a key role in predicting diabetes, but inherited genes are not something that we can easily modify. We should therefore target areas that can be easily improved. Education is key. For example, women can be advised that further pregnancies can increase their risk of diabetes. Furthermore, there is evidence that the strongest risk factor for a child having type 2 diabetes is exposure to diabetes in utero (OR = 10.41, $p<0.0001$).[2] This fact is key to breaking the cycle of diabetes continuing from generation to generation. Education focus should be on reducing glucose levels and BMI scores by encouraging healthier diets with a reduction in sugary and fatty foods and an increase in exercise. If women can reduce their risk of having diabetes, by eating healthier foods and exercising more, particularly during their child bearing years, then it may be possible to reduce the number of children being exposed in utero. Screening for type 2 diabetes may be helpful here as it can be an asymptomatic condition. There still appears to be an information gap regarding why some ethnicities are more prone to diabetes than others, and further research into this area would be beneficial.

Whilst genetic and lifestyle factors are important in diabetes prediction, historical and social factors are also significant factors that have contributed to the prevalence of diabetes in the Pima over the years. In the 1920's, damming of the Gila River resulted in the Pima being displaced from their land and led to a reduction in their traditional low fat, high fibre diet and physical activity levels. Further governmental policies over the years encouraged the Pima towards sedentary jobs and consequently poorer, unhealthy lifestyles.[3] Future studies on diabetes should perhaps pay attention to the broader picture and include social, political and historical factors also.

References

[1] Joseph M. Yracheta et al. "Diabetes and Kidney Disease in American Indians: Potential Role of Sugar-Sweetened Beverages". In: *Mayo Clinic Proceedings* 90.6 (2015), pp. 813–823.

[2] D. Dabelea et al. "Increasing prevalence of Type II diabetes in American Indian children". In: *Diabetologia* 41 (1998), pp. 904–910.

[3] Clayton Booth et al. "Policy and Social Factors Influencing Diabetes among Pima Indians in Arizona, USA". In: *Public Policy and Administration Research* 7.3 (2017), pp. 35–39.

Appendix A: Reference Graphs

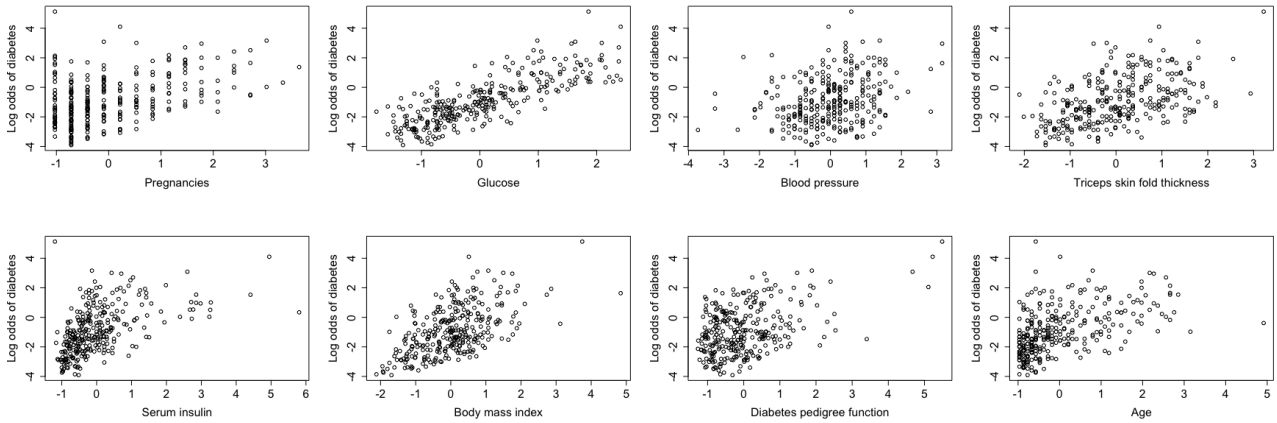

Figure 8: Scatterplots to check linear relationship between the transformed response and the feature variables.

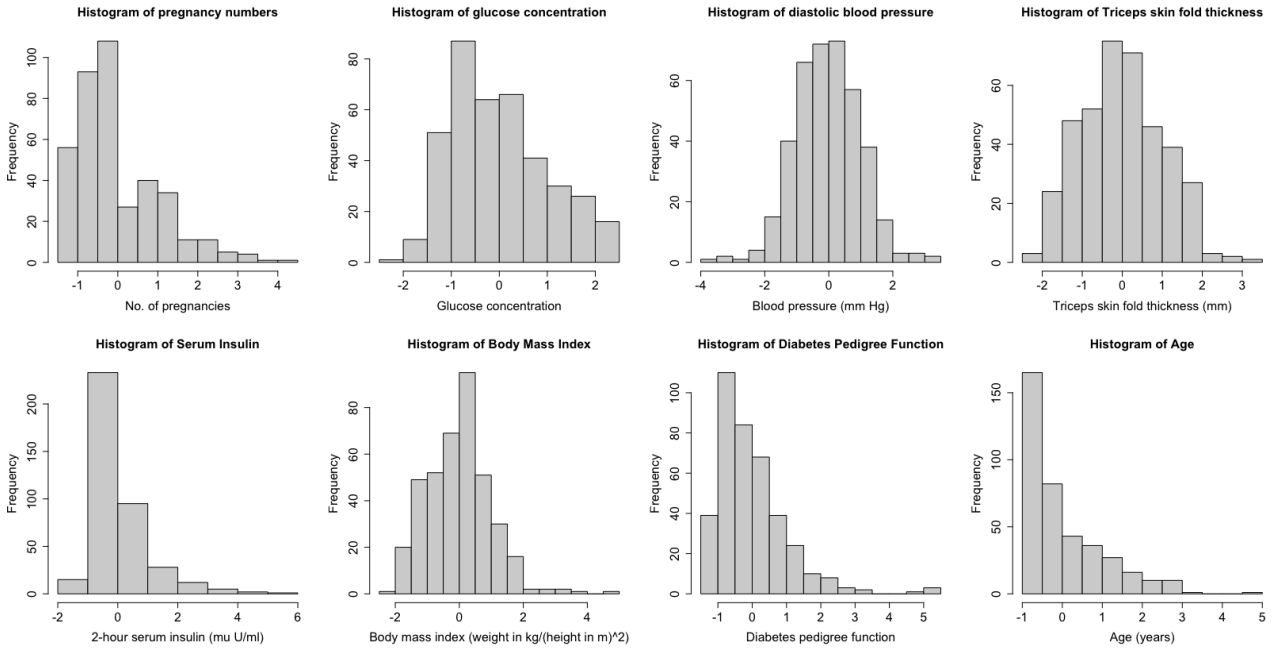

Figure 9: Scaled data histograms

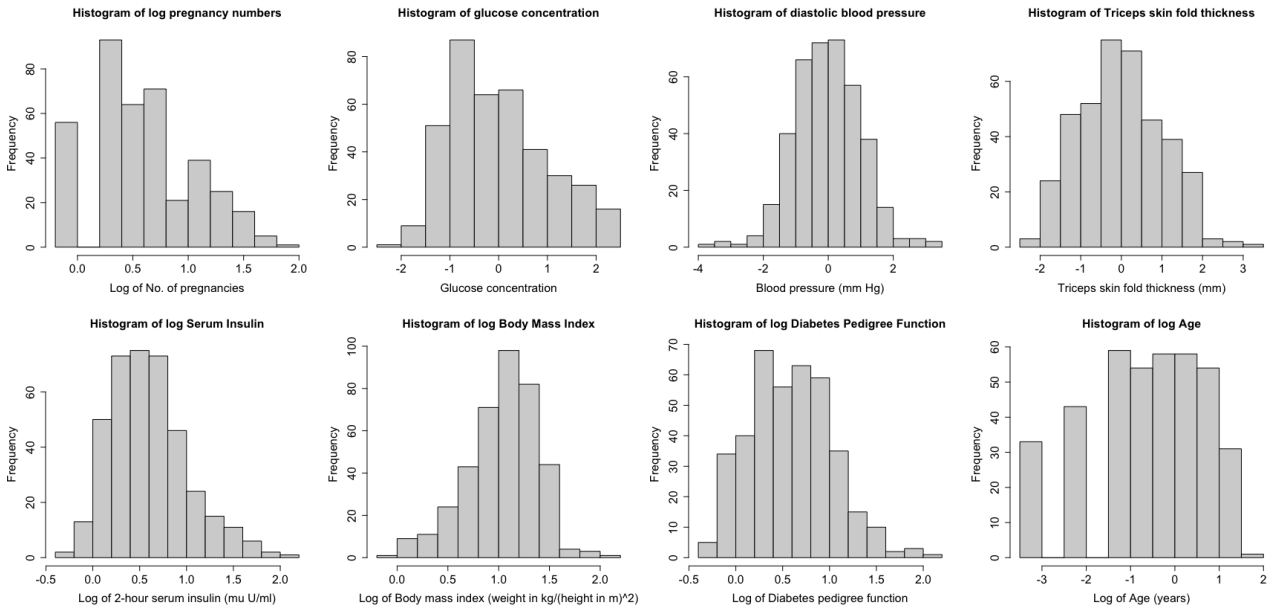

Figure 10: Scaled and transformed data histograms

Appendix B: R code used for analysis

```
#libraries required
library(lattice) #for barchart
library(corrplot) #for correlation plot
library(Epi) #for ROC curve
library(factoextra) #for biplot

# Data
AllData <- read.csv("MVA-Project-Data.csv")
data <- subset(AllData, Student != 202078239 | Student %in% c(NA))

str(data)
data <- data[, 1:17] #keep data we want
str(data)
class(data)
data$diabetes <- as.factor(data$diabetes)
summary(data[,1:9])

#Check if there are any NAs in data
sapply(data, function(x) sum(is.na(x)))

###################### Investigate data ################################

#barplot of response variable
tab <- table(data$diabetes)
# % pos/neg diabetes
prop.table(tab)*100

barchart(table(data$diabetes), horizontal = F, xlab="Diabetes", ylab="No. of Pima women tested",
         col=c("light green", "red"), main="Diabetes test result for Pima women")

#histograms and barplot of data
par(mfrow=c(3,3))
barplot(table(data$pregnant), xlab="No. of pregnancies", ylab="Frequency",
        main="Barchart of number of pregnancies",
        cex.lab=1.5, cex.main=1.5, cex.axis=1.5, ylim=c(0,100))
hist(data$glucose, xlab="Glucose concentration",
```

```
     main="Histogram of glucose concentration", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$pressure, xlab="Blood pressure (mm Hg)",
     main="Histogram of diastolic blood pressure", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$triceps, xlab="Triceps skin fold thickness (mm)",
     main="Histogram of Triceps skin fold thickness",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$insulin, xlab="2-hour serum insulin (mu U/ml)",
     main="Histogram of Serum Insulin", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$mass, xlab="Body mass index (weight in kg/(height in m)^2)",
     main="Histogram of Body Mass Index",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$pedigree, xlab="Diabetes pedigree function",
     main="Histogram of Diabetes Pedigree Function",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$age, xlab="Age (years)",
     main="Histogram of Age",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)

#boxplots of data
par(mfrow=c(2,4))
boxplot(data$pregnant~data$diabetes, ylab="No. of pregnancies",
        main="Boxplot of no. of pregnancies",
        cex.lab=1.5, cex.main=1.5, cex.axis=1.5, xlab="Diabetes test")
boxplot(data$glucose~data$diabetes, ylab="Glucose concentration",
        main="Boxplot of glucose concentration",
        cex.lab=1.5, cex.main=1.5, cex.axis=1.5, xlab="Diabetes test")
boxplot(data$pressure~data$diabetes, ylab="Blood pressure (mm Hg)",
        main="Boxplot of diastolic blood pressure",
        cex.lab=1.5, cex.main=1.5, cex.axis=1.5, xlab="Diabetes test")
boxplot(data$triceps~data$diabetes, ylab="Triceps skin fold thickness (mm)",
        main="Boxplot of Triceps skin fold thickness",
        cex.lab=1.5, cex.main=1.5, cex.axis=1.5, xlab="Diabetes test")
boxplot(data$insulin~data$diabetes, ylab="2-hour serum insulin (mu U/ml)",
        main="Boxplot of Serum Insulin",cex.lab=1.5,
        cex.main=1.5, cex.axis=1.5, xlab="Diabetes test")
boxplot(data$mass~data$diabetes, ylab="Body mass index (weight in kg/(height in m)^2)",
        main="Boxplot of Body Mass Index",
        cex.lab=1.5, cex.main=1.5, cex.axis=1.5, xlab="Diabetes test")
boxplot(data$pedigree~data$diabetes, ylab="Diabetes pedigree function",
        main="Boxplot of Diabetes Pedigree Function",
        cex.lab=1.5, cex.main=1.5, cex.axis=1.5, xlab="Diabetes test")
boxplot(data$age~data$diabetes, ylab="Age (years)",
        main="Boxplot of Age",
        cex.lab=1.5, cex.main=1.5, cex.axis=1.5, xlab="Diabetes test")

## covariances and correlations ##
pairs(data[,1:9])
covariance <- cov(data[,1:8])
correlation <- cor(data[,1:8])
print(covariance, digit=1)
print(correlation, digits=1)
par(mfrow=c(1,1))
corrplot(correlation, method="number", type="upper")

#histograms of scaled data
par(mfrow=c(2,4))
hist(data$pregnant.sc, xlab="No. of pregnancies",
     main="Histogram of pregnancy numbers", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$glucose.sc, xlab="Glucose concentration",
     main="Histogram of glucose concentration", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$pressure.sc, xlab="Blood pressure (mm Hg)",
     main="Histogram of diastolic blood pressure", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$triceps.sc, xlab="Triceps skin fold thickness (mm)",
     main="Histogram of Triceps skin fold thickness",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$insulin.sc, xlab="2-hour serum insulin (mu U/ml)",
```

```
     main="Histogram of Serum Insulin",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$mass.sc, xlab="Body mass index (weight in kg/(height in m)^2)",
     main="Histogram of Body Mass Index",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$pedigree.sc, xlab="Diabetes pedigree function",
     main="Histogram of Diabetes Pedigree Function",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$age.sc, xlab="Age (years)",
     main="Histogram of Age",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)

# transformations of scaled data #
summary(data$pregnant.sc)
data$logpregnant.sc <- log(data$pregnant.sc+2)
hist(data$logpregnant.sc, xlab="Log of No. of pregnancies",
     main="Histogram of log pregnancy numbers", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)

summary(data$insulin.sc)
data$loginsulin.sc <- log(data$insulin.sc+2)
hist(data$loginsulin.sc, xlab="Log of 2-hour serum insulin (mu U/ml)",
     main="Histogram of log Serum Insulin",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)

summary(data$mass.sc)
data$logmass.sc <- log(data$mass.sc+3)
hist(data$logmass.sc, xlab="Log of Body mass index (weight in kg/(height in m)^2)",
     main="Histogram of log Body Mass Index",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)

summary(data$pedigree.sc)
data$logpedigree.sc <- log(data$pedigree.sc+2)
hist(data$logpedigree.sc, xlab="Log of Diabetes pedigree function",
     main="Histogram of log of Diabetes Pedigree Function",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)

summary(data$age.sc)
data$logage.sc <- log(data$age.sc+1)
hist(data$logage.sc, xlab="Log of Age (years)",
     main="Histogram of log Age",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)

# Resulting scaled & transformed data. Don't use though as not much better#
par(mfrow=c(2,4))
hist(data$logpregnant.sc, xlab="Log of No. of pregnancies",
     main="Histogram of log pregnancy numbers", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$glucose.sc, xlab="Glucose concentration",
     main="Histogram of glucose concentration", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$pressure.sc, xlab="Blood pressure (mm Hg)",
     main="Histogram of diastolic blood pressure", cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$triceps.sc, xlab="Triceps skin fold thickness (mm)",
     main="Histogram of Triceps skin fold thickness",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$loginsulin.sc, xlab="Log of 2-hour serum insulin (mu U/ml)",
     main="Histogram of log Serum Insulin",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$logmass.sc, xlab="Log of Body mass index (weight in kg/(height in m)^2)",
     main="Histogram of log Body Mass Index",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$logpedigree.sc, xlab="Log of Diabetes pedigree function",
     main="Histogram of log Diabetes Pedigree Function",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
hist(data$logage.sc, xlab="Log of Age (years)",
     main="Histogram of log Age",cex.lab=1.5, cex.main=1.5, cex.axis=1.5)


# Split data into training and test datasets
set.seed(100)
test.rows <- rbinom(nrow(data), 1, prob=0.25)
train.dat <- subset(data, test.rows==0) # 75% for training set
test.dat <- subset(data, test.rows==1) # 25% for test set
dim(train.dat) #298 obs
dim(test.dat) #93 obs
```

```
# Training Model with all scaled variables (untransformed model)
full.mod1.train <- glm(diabetes=="pos" ~ pregnant.sc + glucose.sc + pressure.sc + triceps.sc +
                         insulin.sc + mass.sc + pedigree.sc + age.sc,
                       data=train.dat, family=binomial)
summary(full.mod1.train)

#transformed model
full.mod2.train <- glm(diabetes=="pos" ~ logpregnant.sc + glucose.sc + pressure.sc + triceps.sc +
                         loginsulin.sc + logmass.sc + logpedigree.sc + logage.sc,
                       data=train.dat, family=binomial)
summary(full.mod2.train)

####### Continued with un-transformed feature variables ##############
# Use full.mod.1.train as full model #
full.mod1.train <- glm(diabetes=="pos" ~ pregnant.sc + glucose.sc + pressure.sc + triceps.sc +
                         insulin.sc + mass.sc + pedigree.sc + age.sc, data=train.dat, family=binomial)
summary(full.mod1.train)
drop1(full.mod1.train, test="Chi")
#######
m <- glm(diabetes=="pos" ~ pregnant.sc + glucose.sc + pressure.sc +
           insulin.sc + mass.sc + pedigree.sc + age.sc, data=train.dat, family=binomial)
drop1(m, test="Chi")
######
m2 <- glm(diabetes=="pos" ~ pregnant.sc + glucose.sc + insulin.sc + mass.sc + pedigree.sc +
            age.sc, data=train.dat, family=binomial)
drop1(m2, test="Chi")
######
m3 <- glm(diabetes=="pos" ~ pregnant.sc + glucose.sc + mass.sc + pedigree.sc + age.sc, data=train.dat,
          family=binomial)
drop1(m3, test="Chi")
######
final.mod.train <- glm(diabetes=="pos" ~ pregnant.sc + glucose.sc + mass.sc + pedigree.sc,
                       data=train.dat, family=binomial)
drop1(final.mod.train, test="Chi")
summary(final.mod.train)

#########################################################################
# Check chi-sqr test for difference in deviances
anova(final.mod.train, full.mod1.train, test="Chi")

#Confidence intervals for model coefficients
cbind(final.mod.train$coefficients, confint.default(final.mod.train)) #log odds CI
exp(cbind(final.mod.train$coefficients, confint.default(final.mod.train))) #odds CI
#########################################################################
#Assess linearity assumption
probs <- predict(final.mod.train, type="response")
logits <- log(probs/(1-probs))
par(mfrow=c(2,4))
plot(train.dat$pregnant.sc, logits, xlab="Pregnancies", ylab="Log odds of diabetes",
     cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
plot(train.dat$glucose.sc, logits, xlab="Glucose", ylab="Log odds of diabetes",
     cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
plot(train.dat$pressure.sc, logits, xlab="Blood pressure", ylab="Log odds of diabetes",
     cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
plot(train.dat$triceps.sc, logits, xlab="Triceps skin fold thickness", ylab="Log odds of diabetes",
     cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
plot(train.dat$insulin.sc, logits, xlab="Serum insulin", ylab="Log odds of diabetes",
     cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
plot(train.dat$mass.sc, logits, xlab="Body mass index", ylab="Log odds of diabetes",
     cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
```

```
plot(train.dat$pedigree.sc, logits, xlab="Diabetes pedigree function", ylab="Log odds of diabetes",
     cex.lab=1.5, cex.main=1.5, cex.axis=1.5)
plot(train.dat$age.sc, logits, xlab="Age", ylab="Log odds of diabetes",
     cex.lab=1.5, cex.main=1.5, cex.axis=1.5)

#####################################################################
#Training predictions of the linear predictor
#get predictions for training data for ROC curve
pred.train <- predict(final.mod.train, newdata=train.dat)

#ROC curve
par(mfrow=c(1,1))
ROC(pred.train, train.dat$diabetes, plot="ROC", main="ROC curve choosing cut-off value for classifier",
    cex.lab=1.5, cex.main=1.5, cex.axis=1.3)
# Gives best cut-off to be -0.881

#Classification matrix/confusion matrix for training data
table(pred.train>-0.881, train.dat$diabetes)

#Predictions for test data
pred.test <- predict(final.mod.train, newdata=test.dat) #get predictions for test data

#Classification matrix for test data
table(pred.test>-0.881, test.dat$diabetes)

########## Principle component analysis ###############
#Use original set of data
data_o <- data[,1:8]
dim(data_o)
#Principle component analysis
pca_output <- prcomp(data_o, scale=T) #scale = T uses correlation for scale
summary(pca_output)
plot(pca_output, main="How many components?")
mtext(side=1, "Principal components")

pca_output #Look at loadings
fviz_pca_biplot(pca_output, col.var="red") #biplot
```