SS9155 - Assignment 2 - 250620601

Ravin Lathigra

2019-01-28

R Packages & Libraries

```
library(corrplot)     #Visualize Correlation between variables
library(kableExtra)   #Style tables
library(tidyverse)    #contains ggplot2,dplyr,tidyr, readr,purr,tibble,stringr,forcats
library(formatR)      #Improve readability of code
library(e1071)        #Functions for latent class analysis, Fourier transform ect.
library(VIM)          #Knn
library(ggfortify)    #Add on to ggplot2 to allow for more plot types
library(Rtsne)        #Dimension reduction classification
library(caret)        #streamlined model development
library(RColorBrewer)#Control colours of visualizations
library(GGally)       #Contains ggpairs plots
library(lmtest)       #Test for linear assumptions
library(MASS)
library(faraway)
```

```
A2 <- pima

str(A2)
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ pregnant : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ glucose  : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ diastolic: int  72 66 64 66 40 74 50 0 70 96 ...
##  $ triceps  : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ insulin  : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ bmi      : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ diabetes : num  0.627 0.351 0.672 0.167 2.288 ...
##  $ age      : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ test     : int  1 0 1 0 1 0 1 0 1 1 ...
```

```
summary(A2)
```

```
##     pregnant         glucose        diastolic         triceps         insulin           bmi
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5   Median :32.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :31.99
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :67.10
##     diabetes           age             test
##  Min.   :0.0780   Min.   :21.00   Min.   :0.000
##  1st Qu.:0.2437   1st Qu.:24.00   1st Qu.:0.000
##  Median :0.3725   Median :29.00   Median :0.000
##  Mean   :0.4719   Mean   :33.24   Mean   :0.349
##  3rd Qu.:0.6262   3rd Qu.:41.00   3rd Qu.:1.000
##  Max.   :2.4200   Max.   :81.00   Max.   :1.000
```

Question 2

Using data sourced from the National Institute of Diabetes and Digestive and Kidney Diseases pertaining to diabetes in Pima Indians we will explore the relationships between the available predictors and the presence of diabetes.

Question 2a

The diagnosis of diabetes is indicative that the body is resistant to insulin produced by the pancreas. Insulin is required to pass glucose to the body's cells, however when the body builds up a resistance to insulin, it can no longer fuel cells and instead leads to an increase build up of glucose in the blood. The goal of our investigations is to model the diagnosis of diabetes for an individual. Within the dataset, diabetes is encoded as a binary response variable with 0 corresponding to as negative diagnosis and 1 otherwise.

If an individual has diabetes, we may expect increased levels of insulin and volatile or increased blood sugar levels. `Figure 1.0` illustrates the distribution of insulin levels split by known diagnosis of diabetes. The plot demonstrates a few interesting observations which seem counterintuitive.

- If a positive diagnosis of diabetes indicates that the body is resistant to insulin, it would be expected that the distribution of insulin of diabetics be shifted towards higher levels than non diebetics. The plot confirms this intuition.
- An intersting set of observation are those with 0 insulin. Humans cannot have a zero insulin level i.e there is a minimum amount required to regulate blood sugar.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

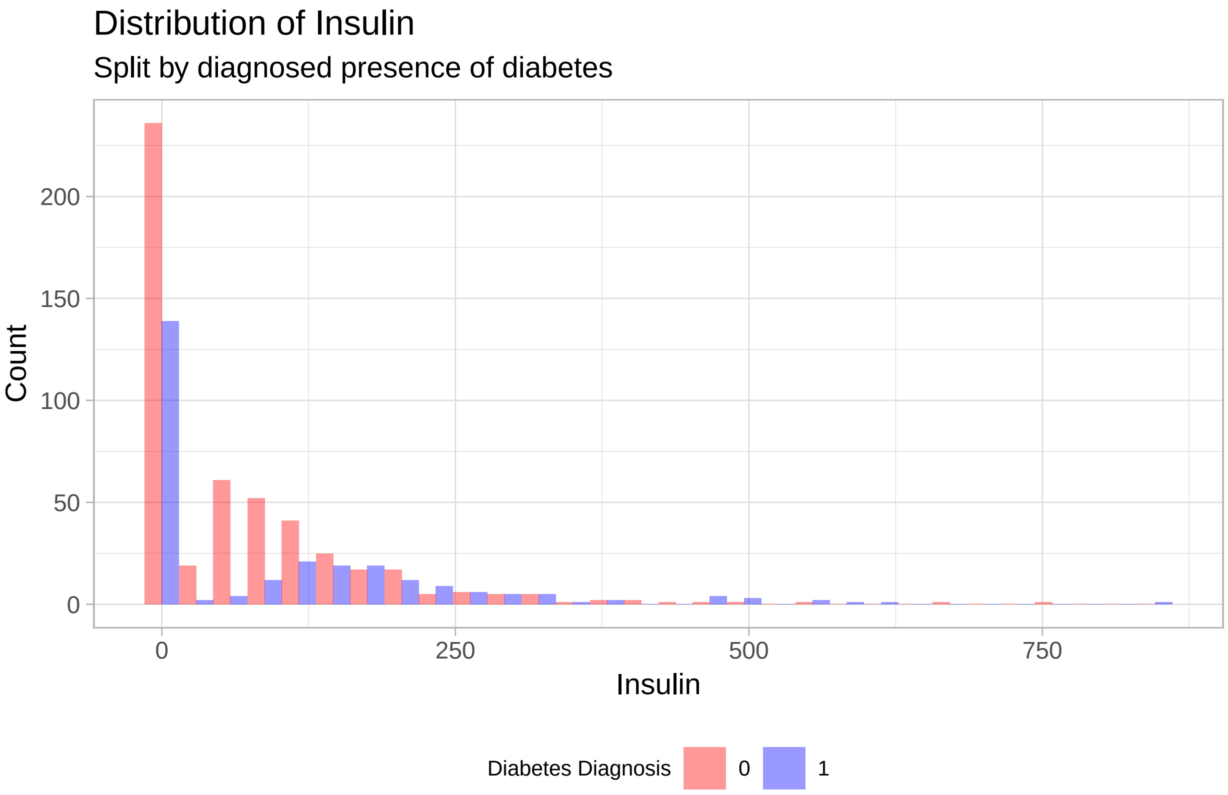


Figure 1.0 | Source: Pima Dataset http://archive.ics.uci.edu/ml/

Question 2b

Zero insulin levels are not valid observations, therefore it would be interesting to observe the distribution of insulin levels without considering these NA observations. `Figure 2.0` illustrates the distribution of insulin levels split by known diagnosis of diabetes while omiting NA values. This plot better represents the differences in distributions of insulin levels between observations from individuals with and without diabetes. As expected, the insulin levels in those diagnosed with diabetes tend to exceed those with negative diagnosis.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

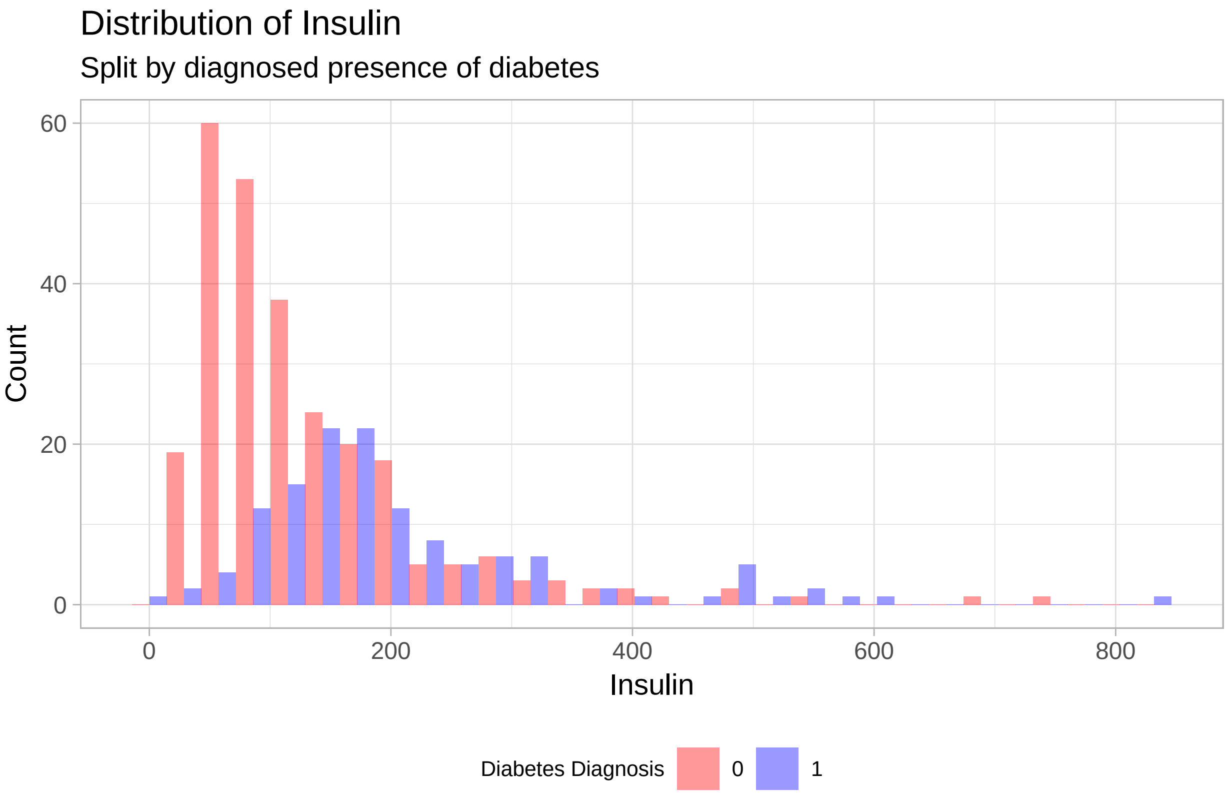


Figure 2.0 | Source: Pima Dataset http://archive.ics.uci.edu/ml/

Question 2c

After replacing NA values from `insulin` we should review other model features that may also require replacing 0's with NA. `Figure 3.0` displays histograms for all features allowing for comparisons between those with and without diabetes. `bmi`, `diastolic`, `glucose`, `pregnant` and `triceps` all have observations that were zero. While `bmi`, `diastolic`, `glucose`, and `triceps` should have strictly positive non-zero values, `pregnant` and indicator for the number of pregnancies that an individual has had may be zero. As such, we will set all zero values to NA for predictors other than `pregnant`.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

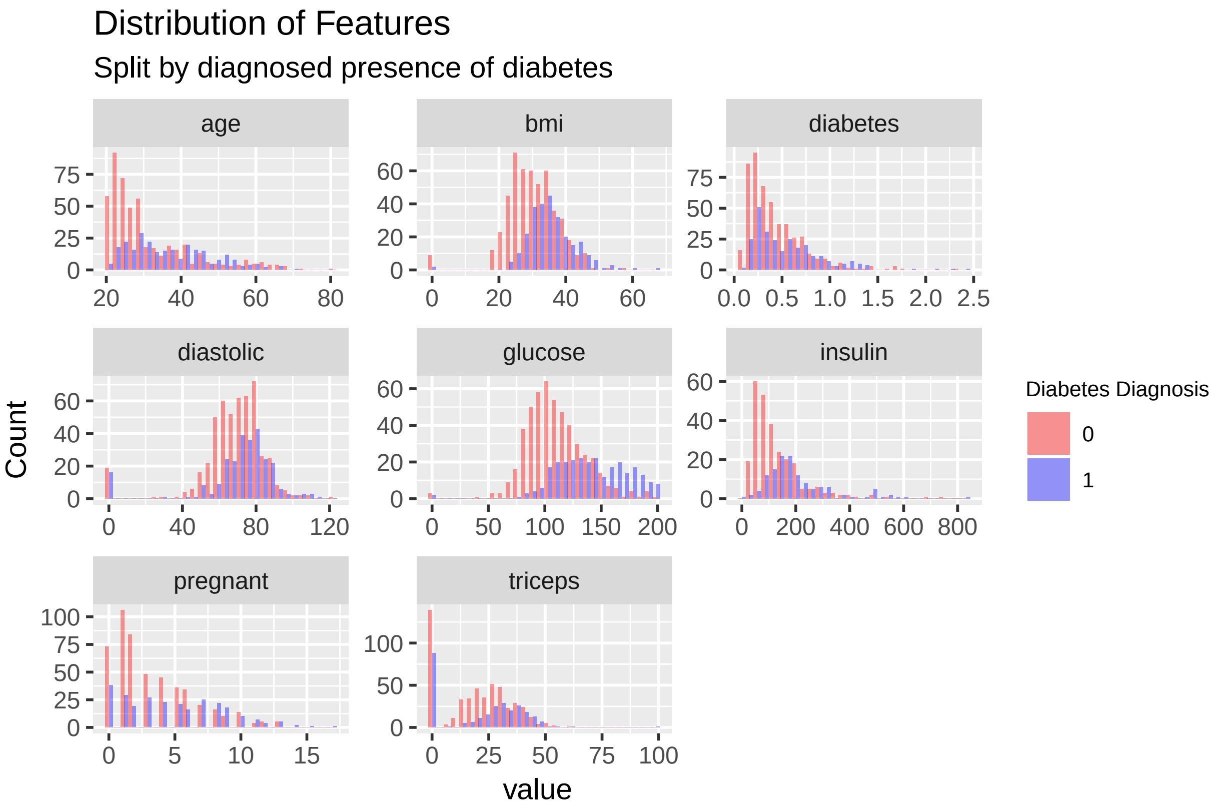


Figure 3.0 | Source: Pima Dataset http://archive.ics.uci.edu/ml/

Once the zero values have been recoded to NA, we can fit a logistic regression tot the data. The following summary output summarizes the final model. Notice that the number of observations used for the model was only 392 compared to the origional 768 records. The reduction in observations is attributable to the observations that have non-NA values across all features.

```
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -1.0041e+01  1.2177e+00 -8.2458 < 2.2e-16
## pregnant     8.2159e-02  5.5426e-02  1.4823  0.138250
## glucose      3.8270e-02  5.7677e-03  6.6351 3.242e-11
## diastolic   -1.4203e-03  1.1833e-02 -0.1200  0.904464
## triceps      1.1221e-02  1.7084e-02  0.6568  0.511279
## insulin     -8.2531e-04  1.3064e-03 -0.6317  0.527565
## bmi          7.0538e-02  2.7342e-02  2.5798  0.009885
## diabetes     1.1409e+00  4.2743e-01  2.6692  0.007603
## age          3.3952e-02  1.8382e-02  1.8470  0.064743
##
## n = 392 p = 9
## Deviance = 344.02123 Null Deviance = 498.09781 (Difference = 154.07657)
```

Question 2d

From the summary provided in the previous question we notice that a few predictors have large p-values. What would happen to our model if we removed `triceps` and `insulin` from the feature space?

The following summary displays the final model considering a reduced feature space:

```
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -9.96047958  1.18187642 -8.4277 < 2.2e-16
## pregnant     0.08404967  0.05507283  1.5262 0.1269713
## glucose      0.03648627  0.00499732  7.3012 2.853e-13
## diastolic   -0.00080021  0.01180337 -0.0678 0.9459488
## bmi          0.07857282  0.02156739  3.6431 0.0002693
## diabetes     1.14923684  0.42503401  2.7039 0.0068537
## age          0.03460789  0.01819187  1.9024 0.0571211
##
## n = 392 p = 7
## Deviance = 344.88054 Null Deviance = 498.09781 (Difference = 153.21727)
```

We can make a direct comparison between these models using analysis of deviance where the null hypothesis is the reduced model and alternative is the full model previously defined. Considering a 5% significance level the following analysis of deviance table yields a p-value of 0.65. Therefore we fail to reject the null hypothesis which leads us to consider removing `triceps` and `insulin` from the feature space.

```
## Analysis of Deviance Table
##
## Model 1: test ~ (pregnant + glucose + diastolic + triceps + insulin +
##     bmi + diabetes + age) - insulin - triceps
## Model 2: test ~ pregnant + glucose + diastolic + triceps + insulin + bmi +
##     diabetes + age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       385     344.88
## 2       383     344.02  2  0.85931   0.6507
```

Question 2e

We have shown that we may consider removing `triceps` and `insulin` from the model, however, we may want to investigate what the best subset of predictors that can be considered are. Using backwards selecting and Akaike Information Criterion, we can identify the best subset of predictors that can be considered. Prior to performing the stepwise feature reduction, we need to omit NA observations.

```
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -9.9920797  1.0868663 -9.1935 < 2.2e-16
## pregnant     0.0839530  0.0550306  1.5256 0.1271173
## glucose      0.0364578  0.0049779  7.3240 2.407e-13
## bmi          0.0781387  0.0206053  3.7922 0.0001493
## diabetes     1.1509128  0.4242424  2.7129 0.0066704
## age          0.0343604  0.0178096  1.9293 0.0536918
##
## n = 392 p = 6
## Deviance = 344.88513 Null Deviance = 498.09781 (Difference = 153.21267)
```

The above summary describes the backward selected model ultimately determining that the predictors most critical to predicting the diagnosis of diabetes are `pregnant`, `glucose`, `bmi`,`diabetes`, and `age`. The model was constructed using 392 observations of which none contained missing values

Question 2f

A drawback we have experienced thus far is that we sacrifice a significant proportion of data to records that had missing values. Perhaps there is a way to preserve this data to some degree to improve model perfomance. By creating a new predictor `check` that is a binary feature indicating the presence of at least one missing record within a particular tuple, we can assess the strength of association between missing records and diagnosis. From there, we can construct a model in which test is the response and `check` is the predictor.

The following analysis of deviance summary, demonstates that there is not a significant relationship between completeness of records and the diagnosis. As such, I suggest that it is most appropriate to consider the stepwise model defined in 2e.

```
## Analysis of Deviance Table
##
## Model 1: test ~ check
## Model 2: test ~ 1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       766     992.43
## 2       767     993.48 -1  -1.0579   0.3037
```

```
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -2.5103540  0.2735029 -9.1785 < 2.2e-16
## pregnant     0.0932305  0.0272208  3.4250 0.0006149
## diabetes     1.1414102  0.2390414  4.7749 1.798e-06
## age          0.0281584  0.0077383  3.6389 0.0002739
##
## n = 768 p = 4
## Deviance = 917.22595 Null Deviance = 993.48391 (Difference = 76.25796)
```

2g

With the final model defined, we can explore relationships of predictors further. In particular, we are interested in the diagnosis of diabetes considering BMI. To explore this, we can explore the following *What is the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant?*

The coefficient of BMI is 0.078.

1st quantile = 28.4 3rd quantile = 37.1

$$Odds = e^{Coef_{BMI}*(quantile_3 - quantile_1)}$$

```
##
## "If all other factors are held constant, the odds of a woman at the third quartile is "
##                                                                                      bmi
##                                                                      "1.9734954858694"
##
##                                " times the odds that of a woman at the first quartile."
```

```
## Waiting for profiling to be done...
```

With the odds calculated, we can develope a 95% confidence interval for this difference. *Table Confidence Interval - BMI* shows the confidence interval for BMI which we can use to develop a confidence interval of the difference.

$$lowerbound = e^{(quantile_3 - quantile_1)*lowerboundBMIConfidenceInterval}$$

$$Upperbound = e^{(quantile_3 - quantile_1)*UpperboundBMIConfidenceInterval}$$

The final confidence interval for the measured difference is shown in the table *95% Confidence Interval - BMI considering differences between 1st and 3rd Quantile*

```
## Waiting for profiling to be done...
```

Table 1: Confidence Interval - BMI

Bounds of Confidence Interval	Value
2.5 %	0.0387490
97.5 %	0.1198444

\begin{table}[!h]

\caption{95% Confidence Interval - BMI considering differences between 1st and 3rd Quantile}

	Lower Bound	Upper Bound
95% Confidence Interval	1.400902	2.836714

\end{table}

2h

To compare the diastolic blood pressure between diagnosis of diabetes, we can utilize box plots. `Figure 4.0` shows the distribution of diastolic blood pressures in women with and without diabetes. It may suggest that the blood pressures of diabetic patients tends to be higher than those with negative diagnosis.

The question is then if we have evidence of eleveated blood pressure in those with postive diagnosis, is this not a significant predictor to consider.

If we refer back to the final model that used `pregnant`, `glucose`, `bmi`,`diabetes`, and `age` as its features, which were selected using backwards selection, why was `diastolic` not considered. This discrepency is explained by, in the presence of the other model predictors, `diastolic` does not contribute significantly to the diagnosis.

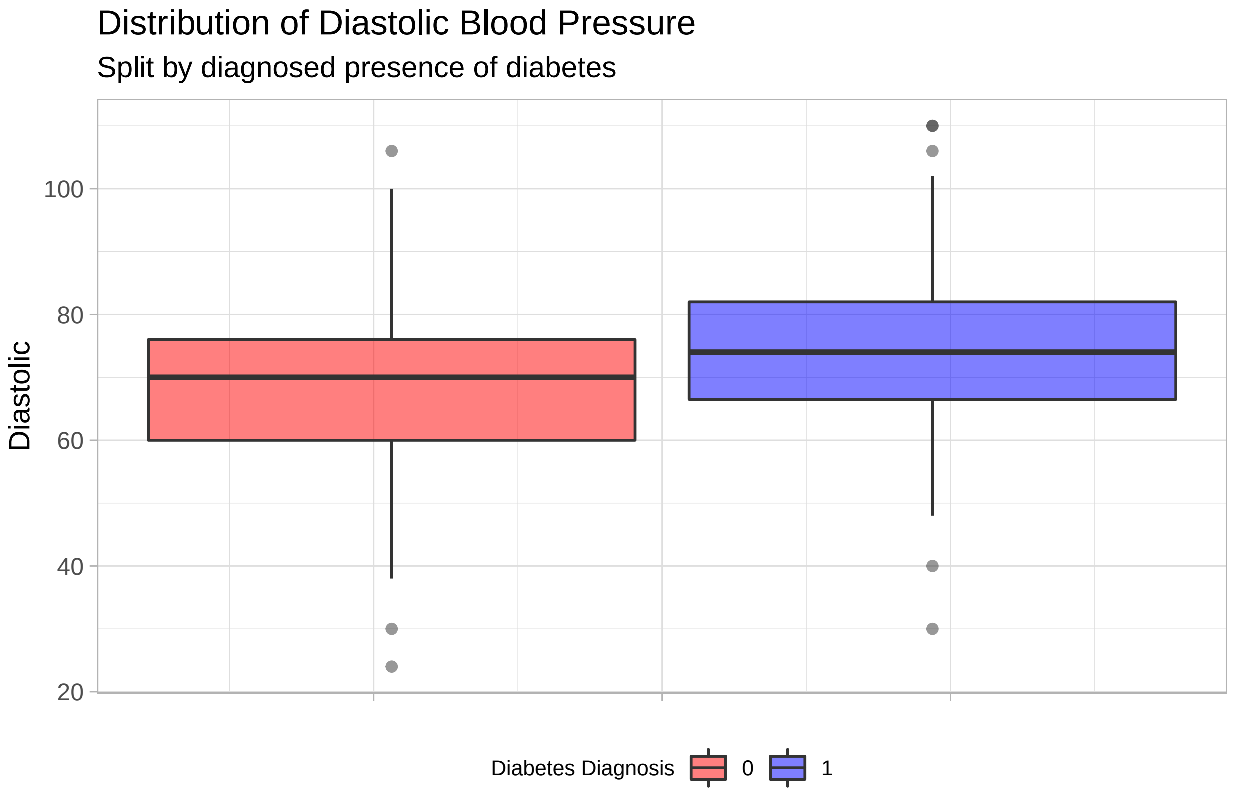


Figure 4.0 | Source: Pima Dataset http://archive.ics.uci.edu/ml/

Question 3

Question 3a

Figures 5.0 - 7.0 illustrate the relationship between the response variable `kyphosis` and the predictors `Age`, `Number`, and `Start`.

+Some immediate observations that can be made from the plots are that age, while there may be a slight tendency for lower ages to to not have kyphosis compared ot those that do as well as decreased variance in the ages that had kyphosis.

+The number of vertebrae involved seems to also tend to be lower for those without kyphosis comared to those that do, though the number of top vertebae operated on seems to be greater for those without kyphosis.

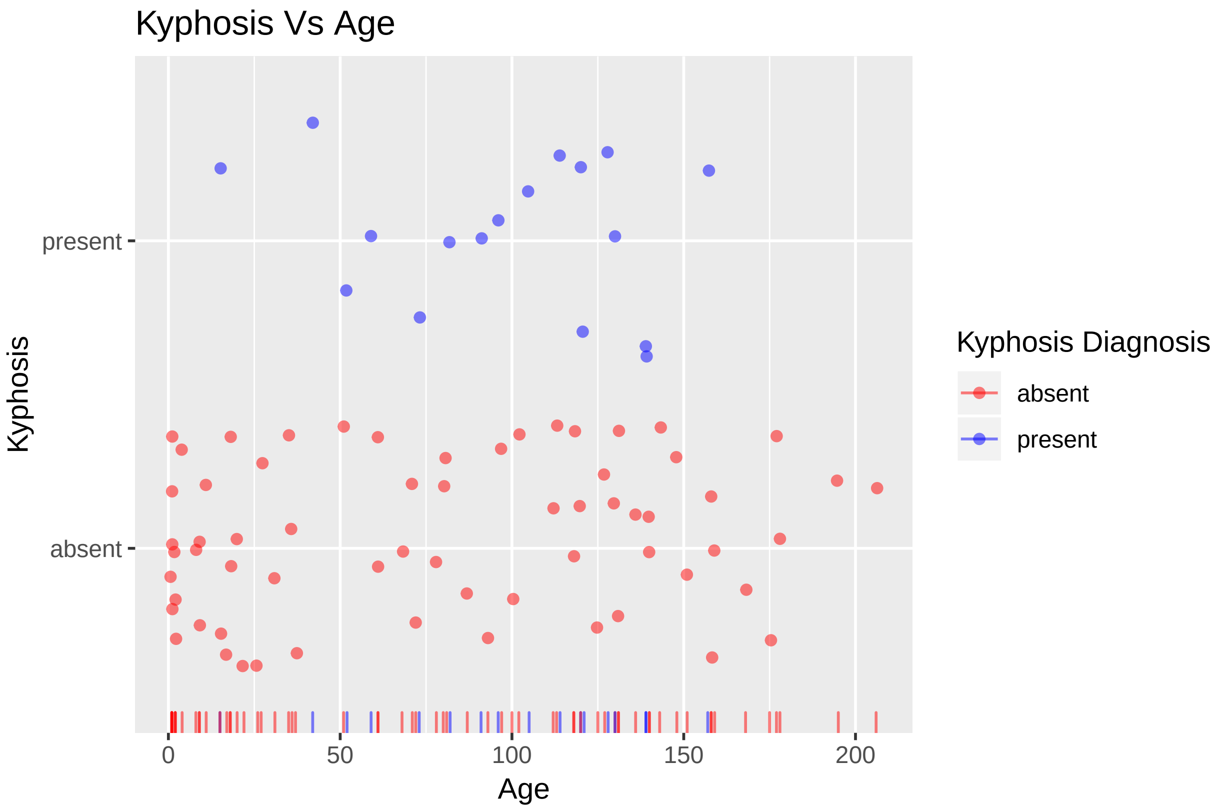


Figure 5.0

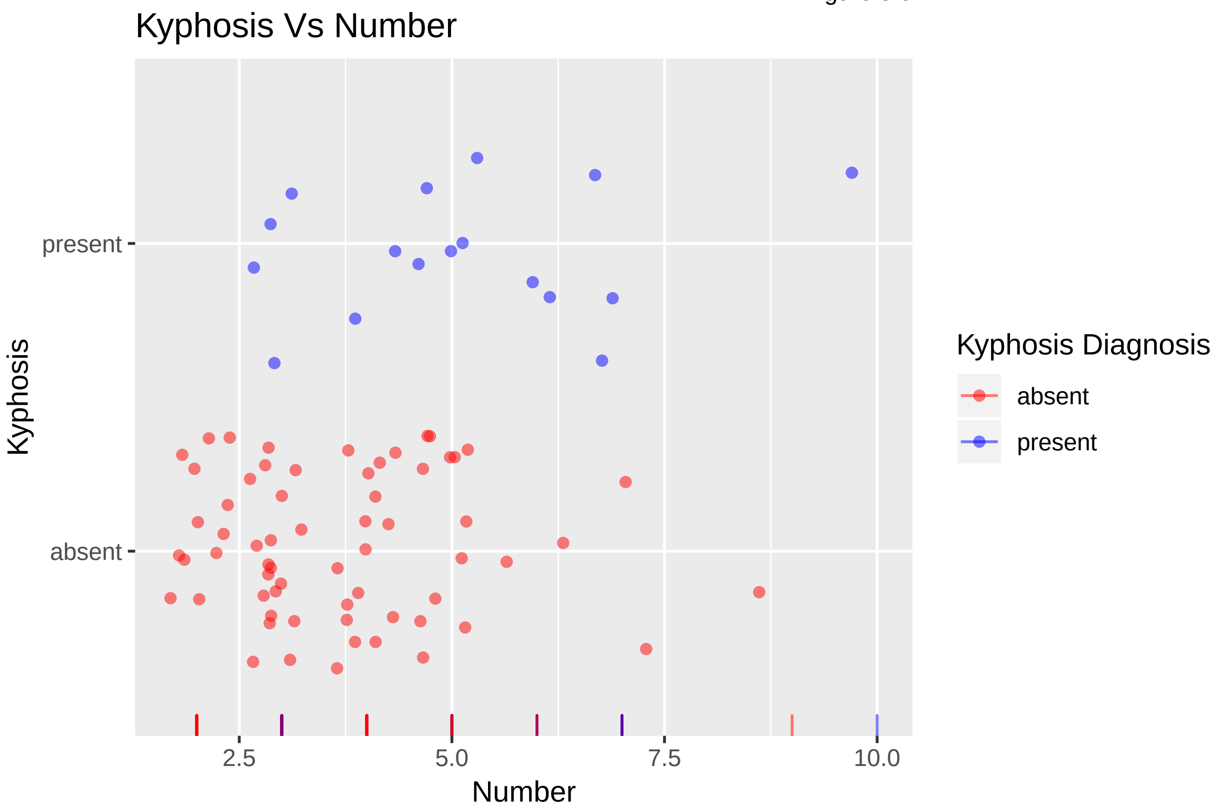


Figure 6.0

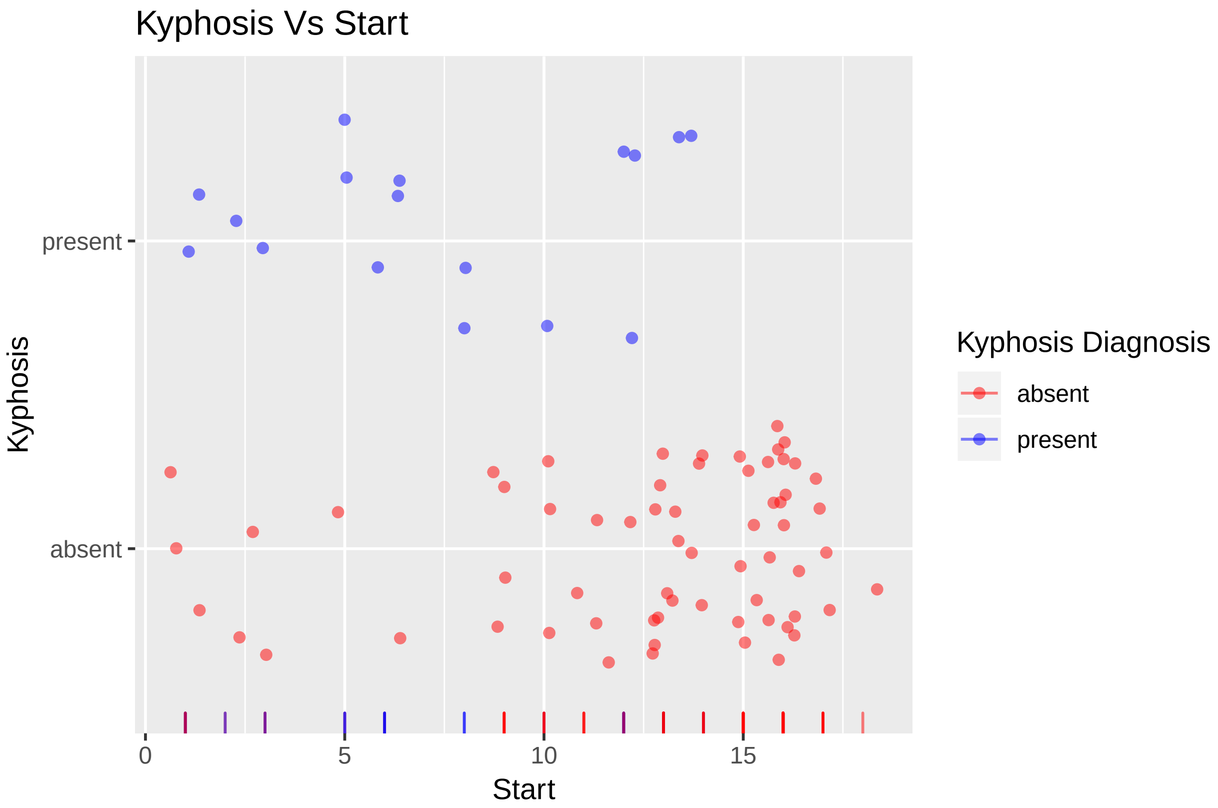


Figure 7.0

3b

The first thing we can do is create a logistic regression where `Kyphosis` depends on `Age`, `Number`, and `Start`.

The following summary outlines the fitted model:

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.0369335  1.4495745 -1.4052 0.159964
## Age          0.0109305  0.0064463  1.6956 0.089955
## Number       0.4106012  0.2248608  1.8260 0.067847
## Start       -0.2065100  0.0676989 -3.0504 0.002285
##
## n = 81 p = 4
## Deviance = 61.37993 Null Deviance = 83.23447 (Difference = 21.85455)
```

To gain further insight, we can visualize the residuals vs the fitted values. As default, the calculated residuals of a logistic regression are deviance residuals. `Figure 8.0` displays the residuals vs fitted values, notice that this plot, is not very helpful. The residual can take only two values given a fixed linear predictor. In this case, the upper line in the plot corresponds to positive diagnosis and the lower, negative.

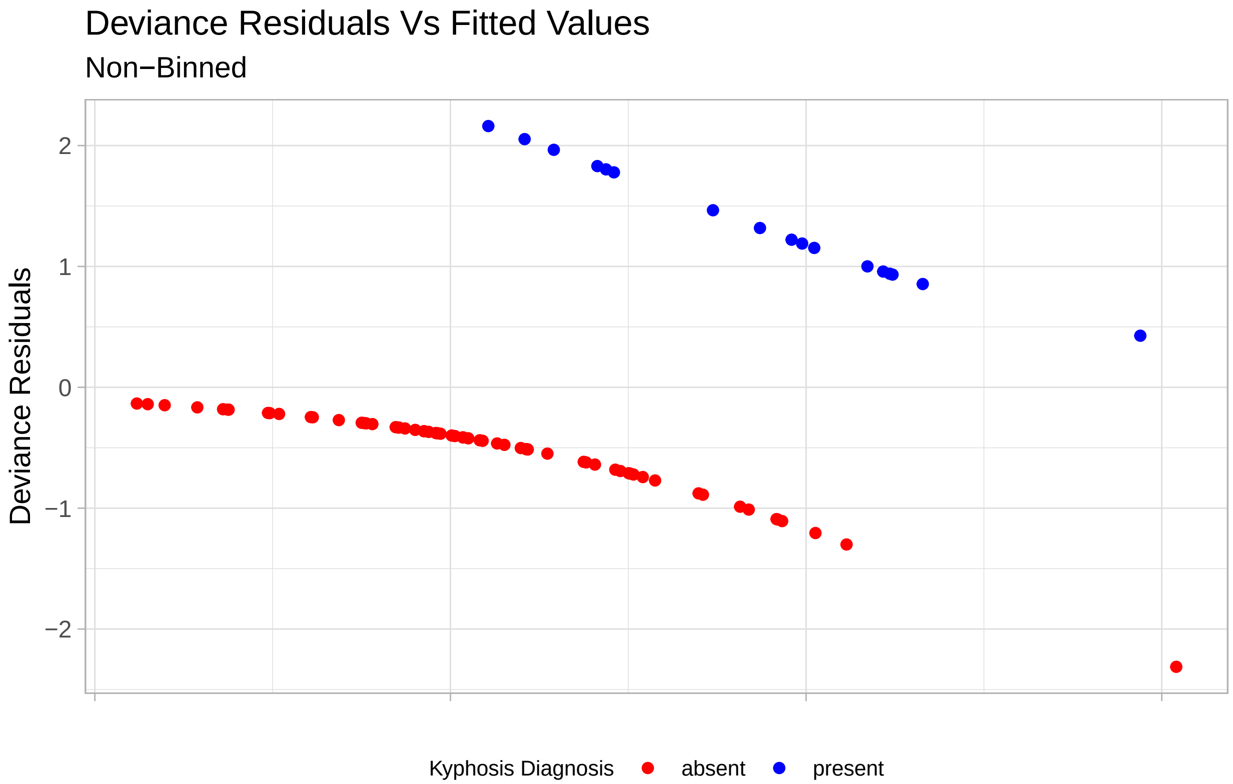
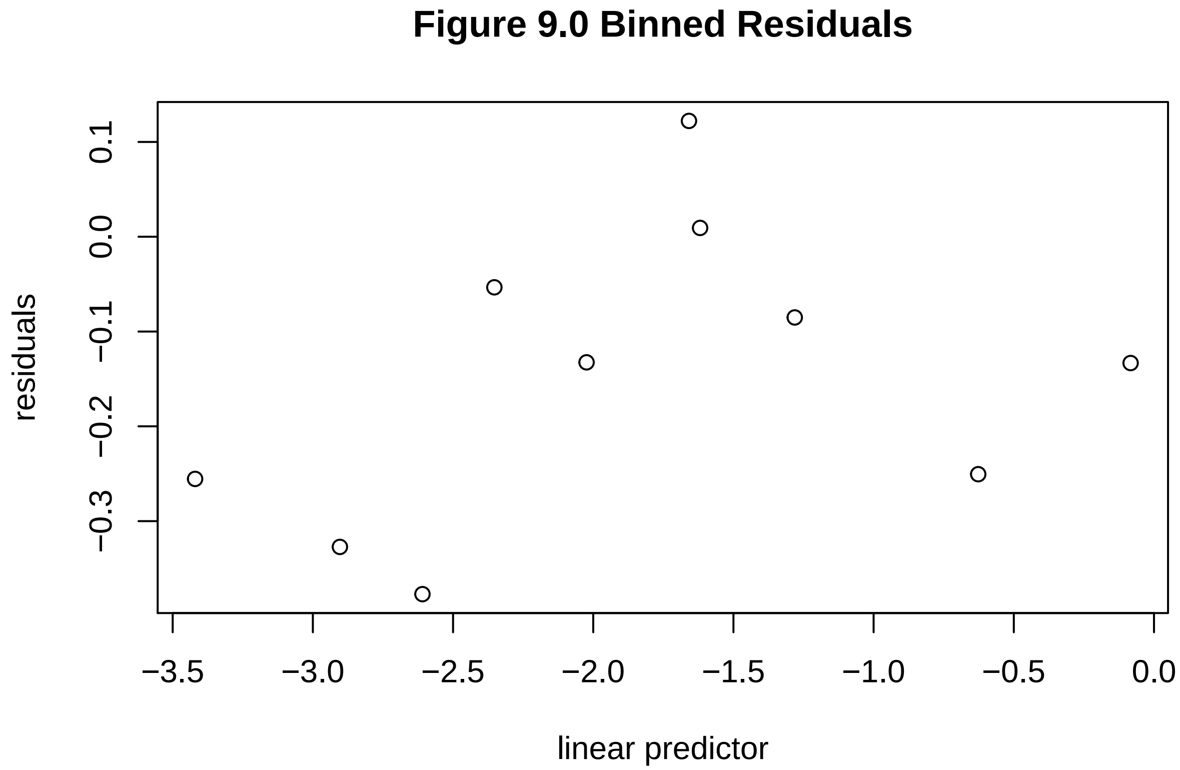


Figure 8.0 | Source: John M. Chambers and Trevor J. Hastie eds. (1992)

3c

Produce a binned residual plot as described in the text. You will need to select an appropriate amount of binning. Comment on the plot.

To add more values to our plots we can bin the residuals. Figure 9.0 plots the binned residuals using 10 bins. Though the number of points is fairly sparse, it seems there is generally non-equal variance. Homoscedacitiy is not required therefore this is not a concern.



To add more values to our plots we can bin the residuals.

3d

Now we want to inspect residuals and the relationship with `Start`. Figure 10 plots residuals vs `Start` it is most important to note the points of the most extreme sizes i.e small and large. Note:We take square roots because the SD is proportional to the square root of the sample size so this gives the appropriate visual impression. In particular most observations had a large number of start top vertebae and the observaitons with more extreme residuals occur with low frequency.

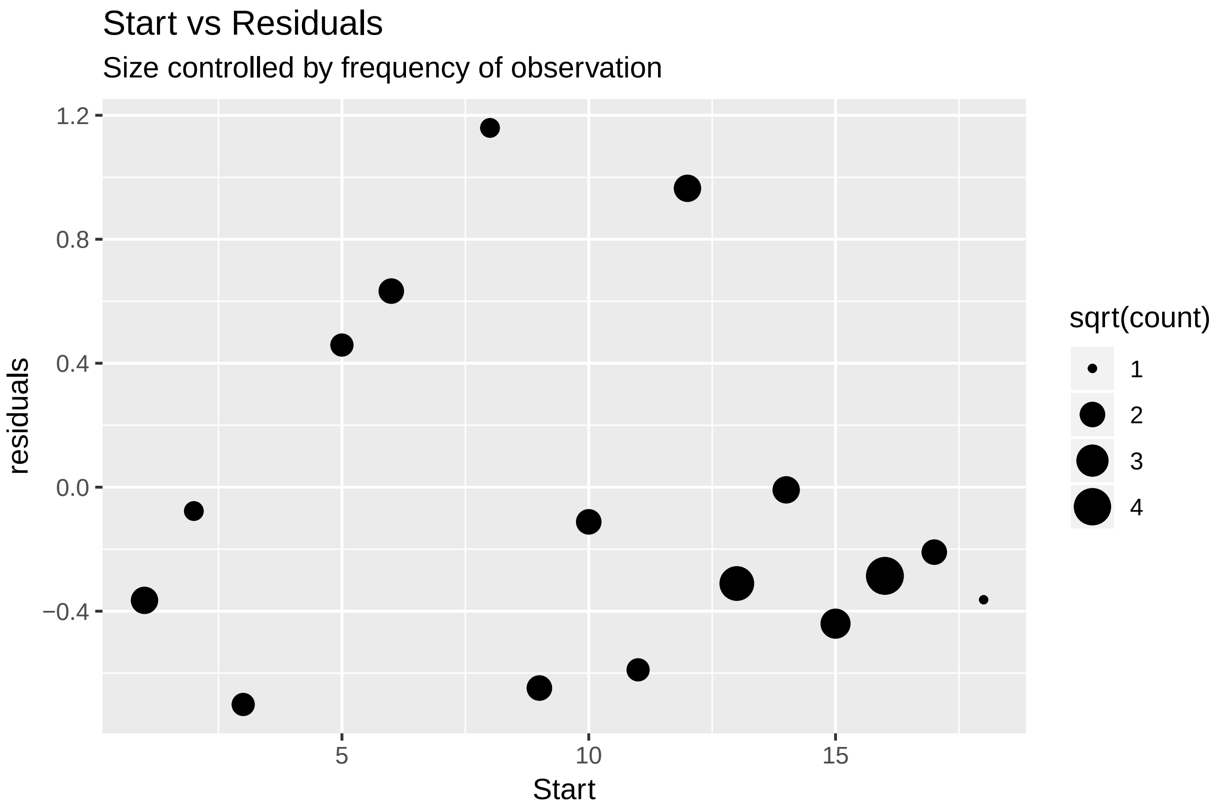
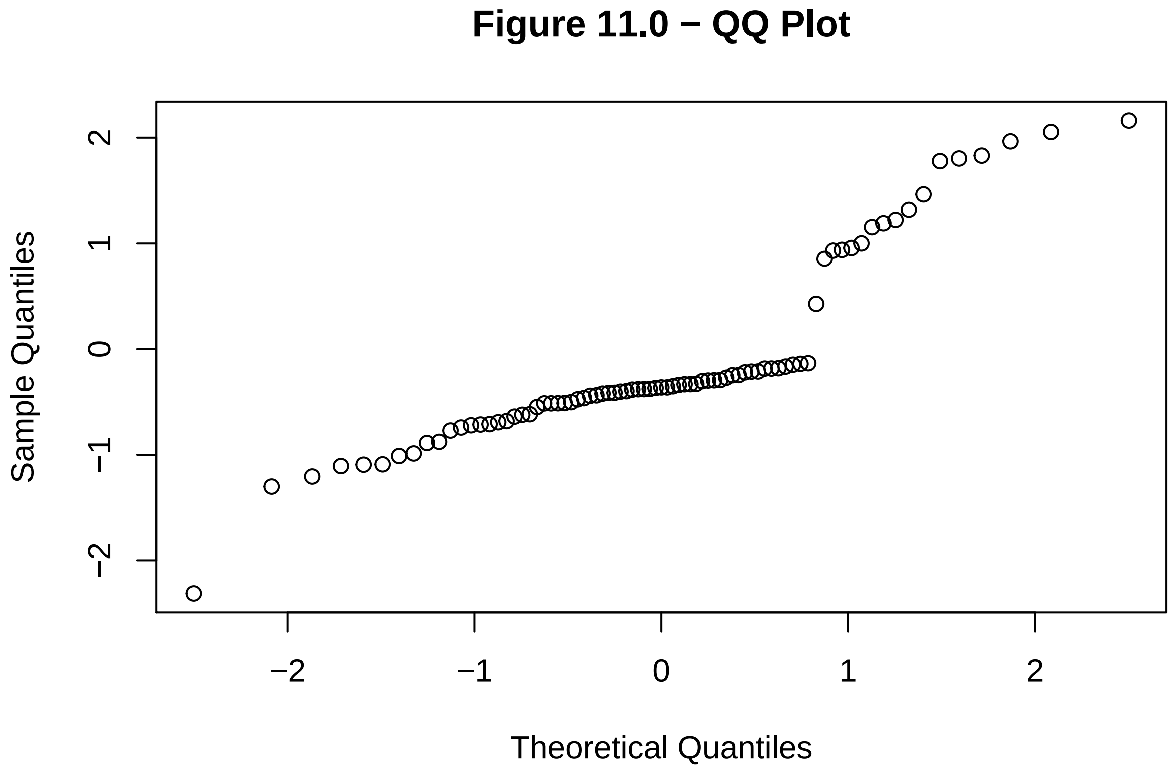


Figure 10.0

3e

To further analyze the model residuals we can examine a qqplot. The residuals need not be normal, so little information can be further gathered from this plot, instead we will need to observe the leverage.

Note: the largest residuals will arise when there was a positive classification with low predicted probability.



3f

As previously mentioned, to gain further insight into our model we need to observe the leverage of our observations. Using the `halfnorm` function in R, we can visualize the points with greatest leverage.

In our case, there were 2 observations with particularly large leverage. They are labelled in the following plot. Futher exploring these points leads us to notice that they correspond to the most cases with the most extreme number of vertebrae involved. These points while they have the largest leverage do not seem to be extreme enough to consider ommision.

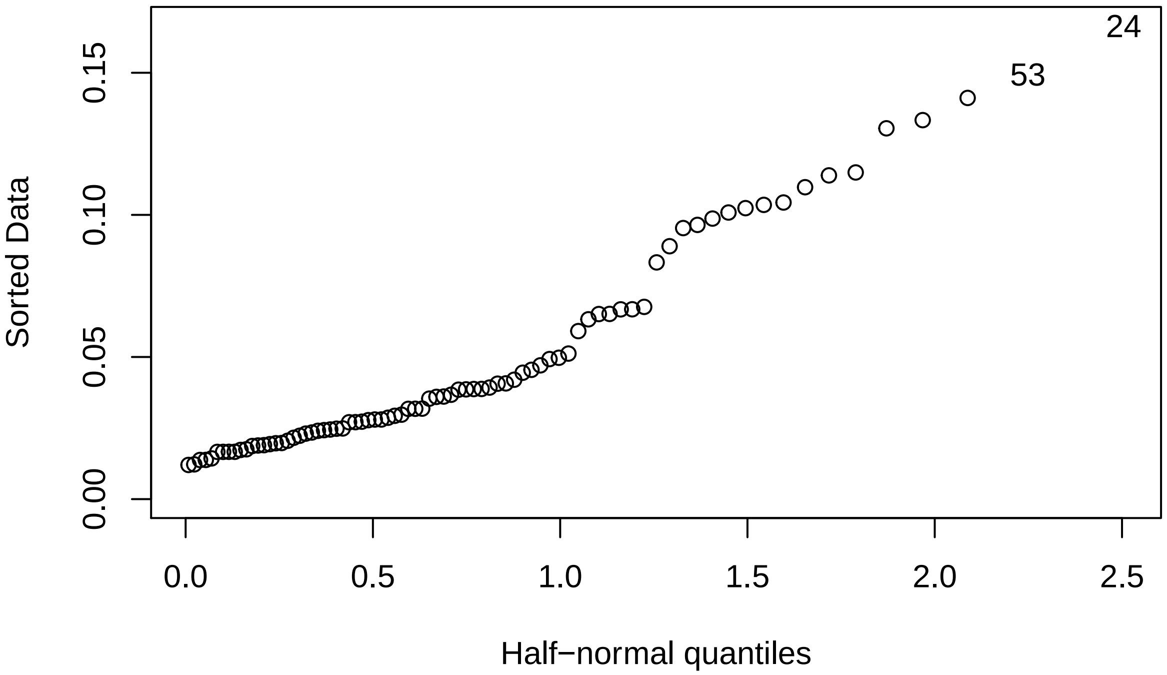


Figure 11.0 − Investigating Leverage

3g

The following plot displays observed proportions vs predicted probability. Although we can see there is some variation, there is no consistent deviation from what is expected and the line passes through most of these intervals which suggests that the variation from the expected is within our comfort level.

3h

With our model complete, we can classify out observations and assess the model accuracy. The following output displays a confusion matrix of predicted vs actual classes. In our case, we were able to predict the absent class well however we were less successful in the positive class. When the patient had Kyphosis, we were only able to predict with 41% accuracy. This could be adjusted by adjusting the prediction threshold downward allowing for more missclassified absent observations.

To improve our model an additional test set to access the unbiased generalization error could provide a better estimate of the true error.

```
##          predout
## Kyphosis  no yes
##   absent  36  28
##   present  1  16
```