

Samuel O'Mahony
19236719

MS4034 – Project 2 Report



28/02/21

Samuel O'Mahony
I.D: 19236719

Introduction

The data I examined in this project details 8 medical variables and a risk score recorded for a sample of 320 female Pima Indians in the USA. The sample I used is a subsample of the data collected as the original dataset was missing data on serum insulin.

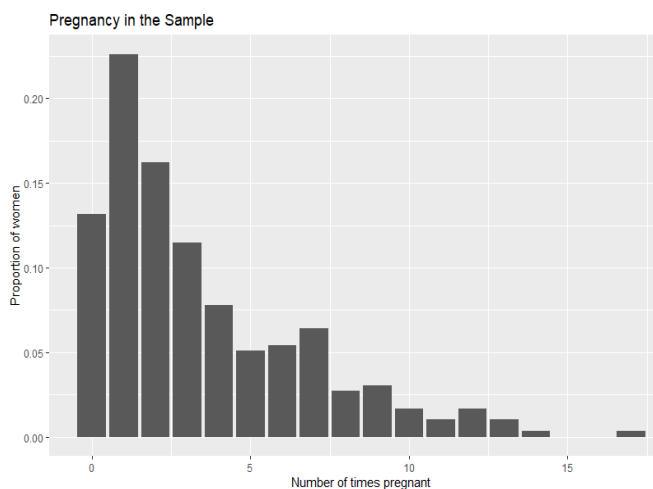
In the following report, I explore each variable in the dataset, providing the necessary characteristics and visualisations to effectively summarise the variables. The distributions of plasma glucose concentration, blood pressure, and serum insulin in the diabetic portion of our sample are compared with the respective distributions in the non-diabetic portion of the sample. The glucose concentration is further examined, and I test and discuss whether there is a statistically significant difference in the mean glucose concentration of those with diabetes and those without. Finally I investigate the data in terms of model fitting and develop a model to predict the risk score of an individual.

Summary of Variables

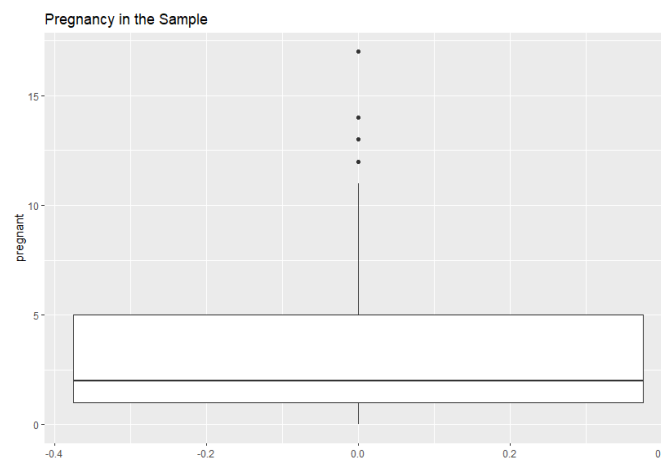
Pregnant

This variable refers to the number of times the woman in question has been pregnant.

Mean	3.396875
Standard Deviation	3.272665
95% Confidence Interval	3.038304, 3.755446
Median	2
Q1 and Q3	1, 5
IQR	4



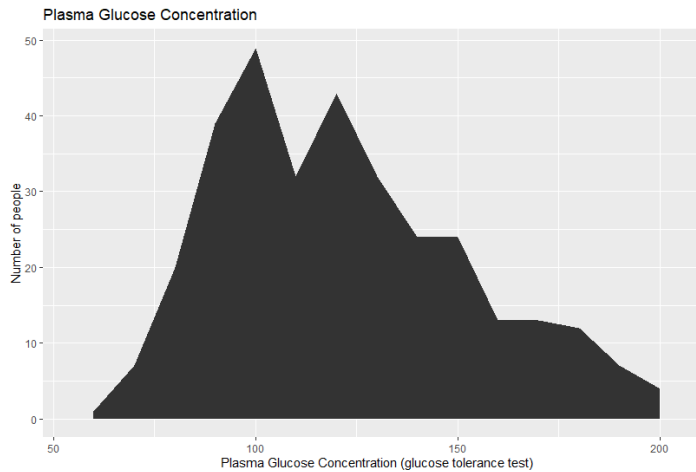
As seen, the data is quite positively skewed. The median seems to be the better measure of centrality here. Given how small the confidence interval is, it does not provide a good understanding of the spread of the data.



There are some outliers that contribute heavily to the skewness of this variable, such as one woman who was pregnant 17 times.

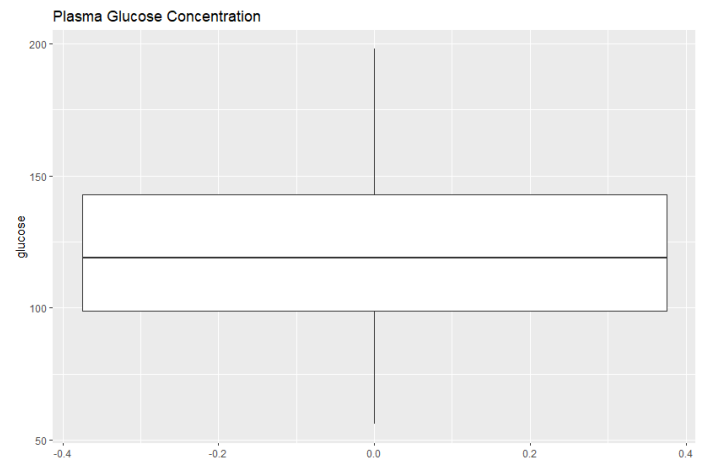
Glucose

The plasma glucose concentration was measured by a glucose tolerance test.



Once again, we notice a positive skew. The median seems to be the better measure of centrality here.

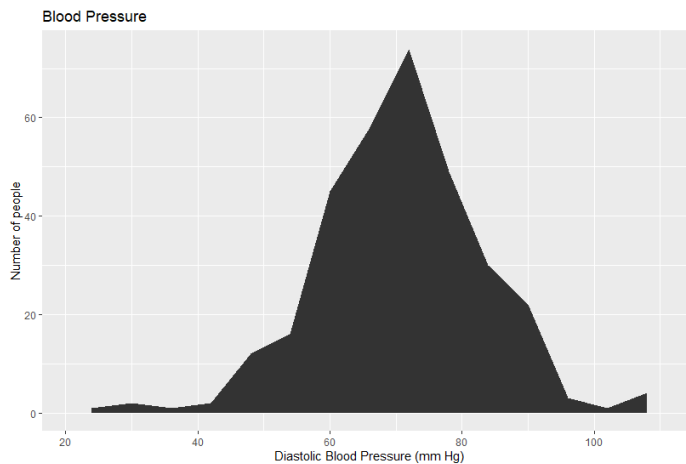
Mean	122.175
Standard Deviation	30.59679
95% Confidence Interval	118.8227, 125.5273
Median	119
Q1 and Q3	99, 143
IQR	44



As seen in the above boxplot, no values lie outside the whiskers. Glucose did not contain many outliers.

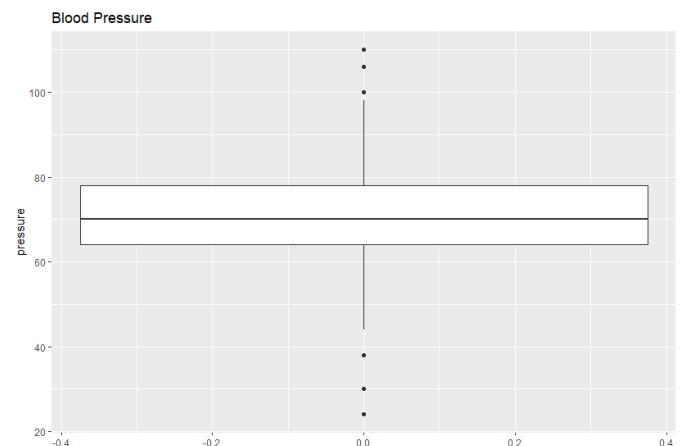
Pressure

The diastolic blood pressure, measured in mm Hg.



We observe slightly negatively skewed data in this case. The mean and median are so close in value that they could almost be used interchangeably but the confidence interval is too small to describe the spread of the data sufficiently.

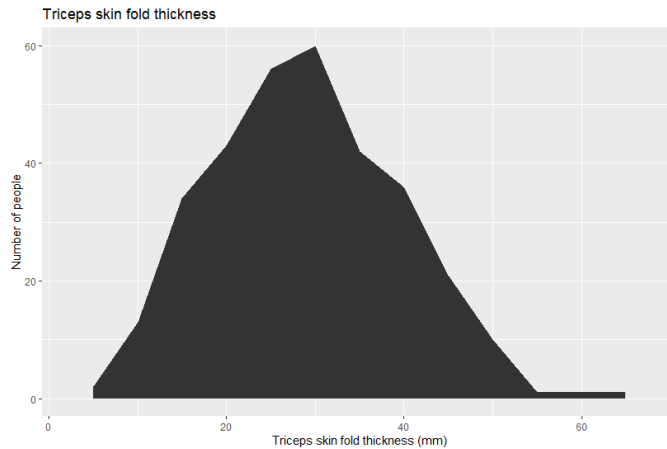
Mean	70.63438
Standard Deviation	12.38543
95% Confidence Interval	69.27736, 71.99139
Median	70
Q1 and Q3	64, 78
IQR	14



Pressure contained some outliers. A few women had quite low blood pressure (<45 mm Hg)

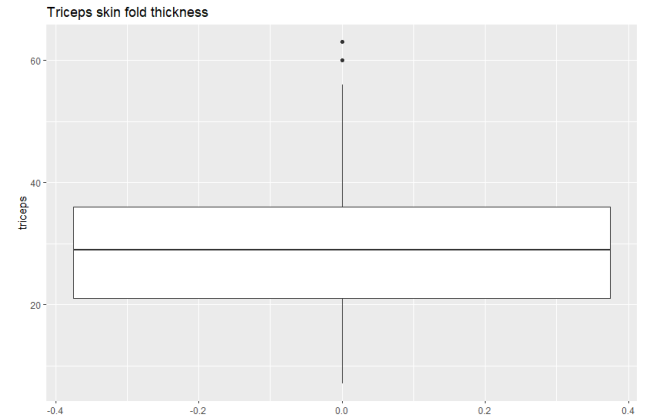
Triceps

This variable describes the skin fold thickness of the triceps. (measured in mm)



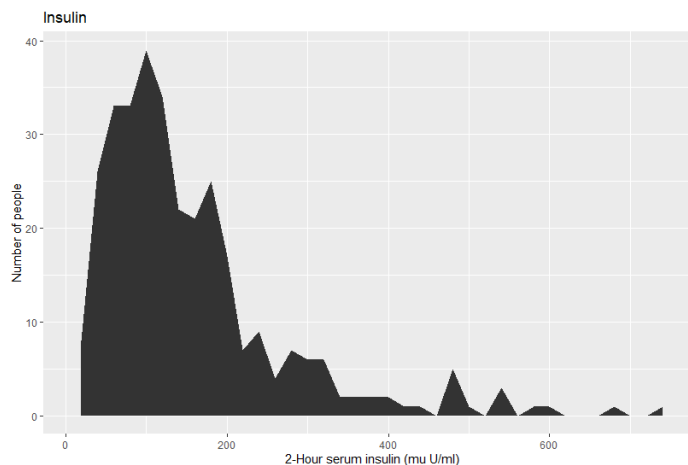
The data is very slightly positively skewed and contains few outliers. This variable and the mass variable have a moderate correlation ($r = 0.66$ using Pearson) and this will affect how we treat this variable later on in forming our model

Mean	28.90938
Standard Deviation	10.56201
95% Confidence Interval	27.75215, 30.06660
Median	29
Q1 and Q3	21, 36
IQR	15



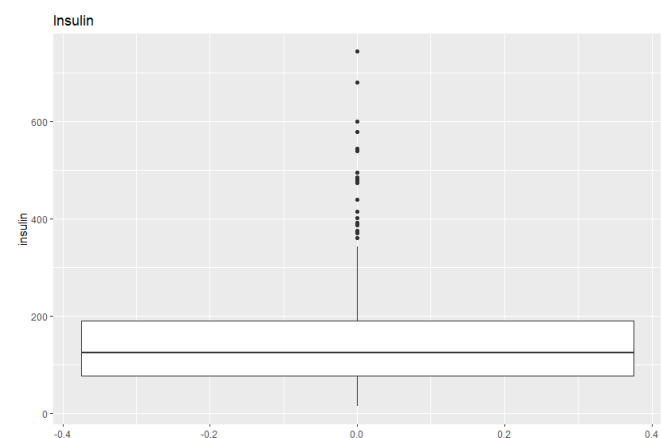
Insulin

A measure of the 2-Hour serum insulin present. (measured in $\mu\text{U/ml}$)



This data is heavily positively skewed. Since there are not many values around zero, there does not seem to be many type I diabetics in the sample. This is hypothesis which I do not test in this report, however.

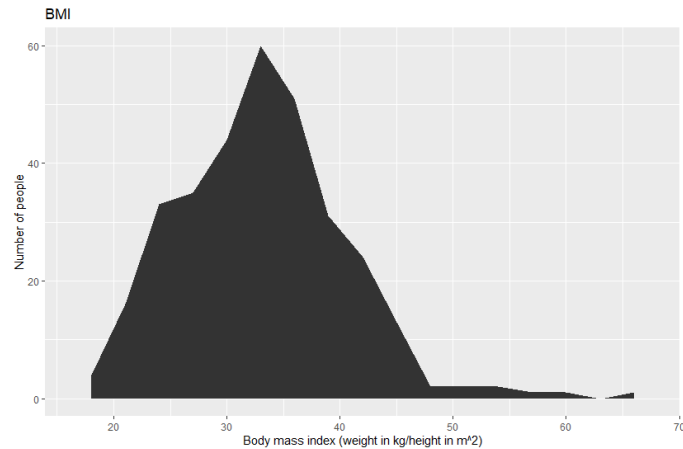
Mean	155.6031
Standard Deviation	116.6135
95% Confidence Interval	142.8263, 168.3799
Median	125
Q1 and Q3	76.75, 190
IQR	113.25



As seen above, there are many outliers that fall far above a normal level of insulin.

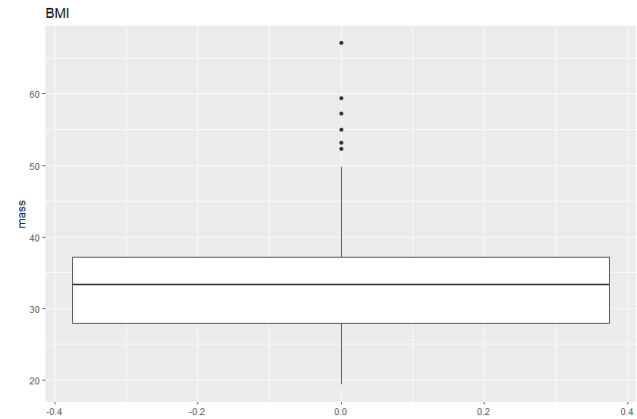
Mass

The Body Mass Index or BMI of the people being sampled.



Once again, we have positively skewed data. Given that our Q1 is 27.8 and according to the NHS, the ideal BMI for most adults is between 18.5 and 24.9, the BMI in our sample is certainly above average. [\[1\]](#)

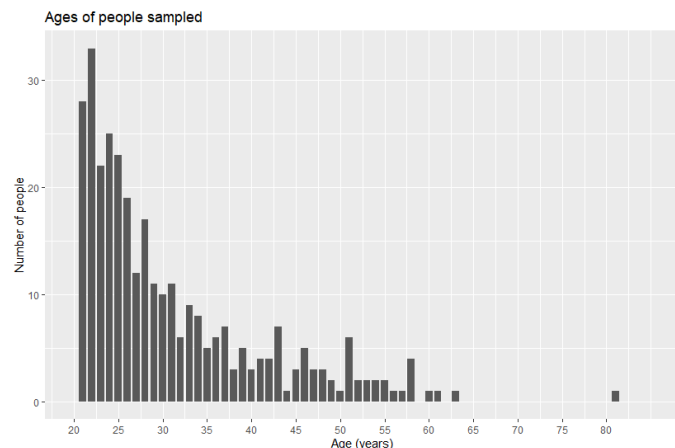
Mean	33.12812
Standard Deviation	7.274369
95% Confidence Interval	32.33111, 33.92514
Median	33.3
Q1 and Q3	27.98, 37.12
IQR	8.7



BMI contains a couple of outliers, most of which are at the higher end of the values recorded. There are not enough outliers that they alone would explain the skewness.

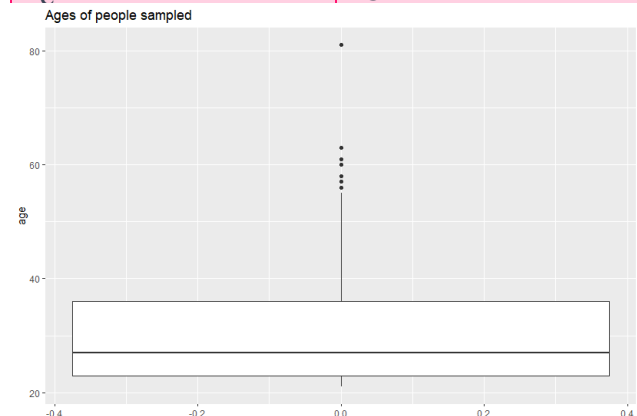
Age

The age of the women who were sampled. It is worth noting that all women included were at least 21, hence the lack of values below 21.



The data is heavily positively skewed, both due to several outliers who have achieved relatively long lives and due to an abundance of women in the 20-25 range.

Mean	30.9625
Standard Deviation	10.24367
95% Confidence Interval	29.84015, 32.08485
Median	27
Q1 and Q3	23, 36
IQR	13



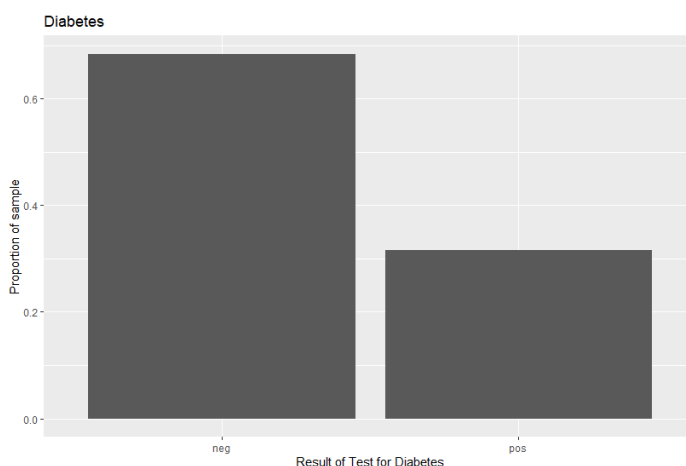
It might be interesting to investigate the factors behind this variable i.e., whether the relatively small number of older women in the community is due to women leaving the community or due to health-related factors lowering the life expectancy, but we do not address this question in this report.

Diabetes

Describes whether or not the woman in question tested positive or negative for diabetes by WHO criteria.

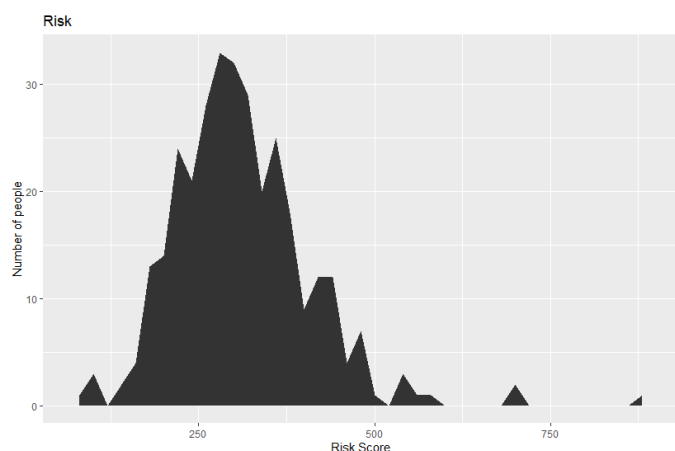
	<i>Frequency</i>	<i>Percent</i>
Negative	219	68.4
Positive	101	31.6
Total	320	100

As seen, approximately two thirds of the population tested negative. Based on the trends seen in the insulin and mass variables, I suspect that the diabetics in the sample are primarily type 2 diabetics, however this is a hypothesis that will remain untested in this report.



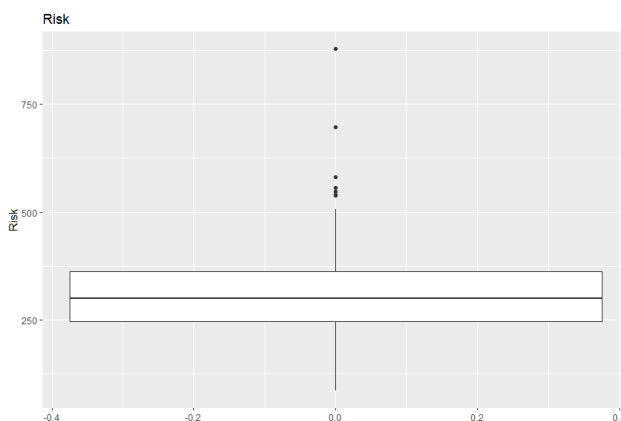
Risk

This refers to a risk score given to each woman sampled that denotes the likelihood of their death over the next 10-year period.



The data here is positively skewed. We see the majority of the risk scores in one cluster (around ~300) but there are a significant number of risk scores far above that cluster.

<i>Mean</i>	311.2093
<i>Standard Deviation</i>	96.24091
<i>95% Confidence Interval</i>	300.6647, 321.7540
<i>Median</i>	299.31
<i>Q1 and Q3</i>	247.05, 361.41
<i>IQR</i>	114.36



Those outliers that are far above the mean seem to majorly contribute to the skew of this data.

Comparison of Glucose, Pressure, and Insulin in those with and without Diabetes

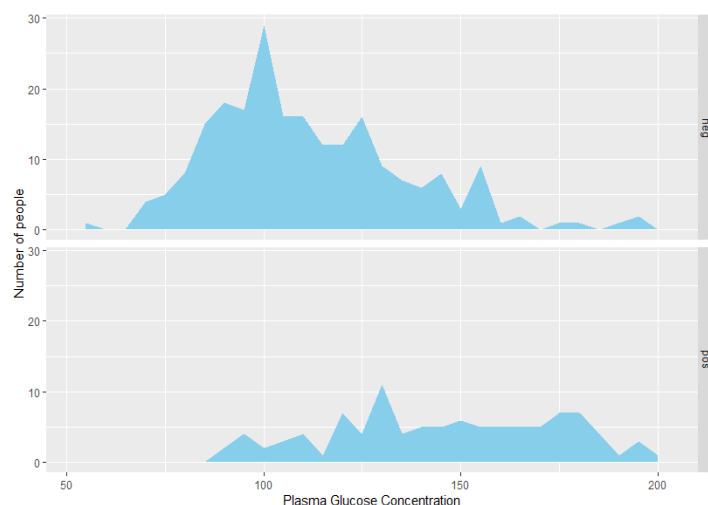
Glucose in those with and Without Diabetes

Glucose	Diabetics	Nondiabetics
Mean	143.6296	112.6085
Standard Deviation	30.48536	24.07714
Median	144.0	109.0
Q1 and Q3	121.8, 171.2	95, 127
IQR	49.4	32

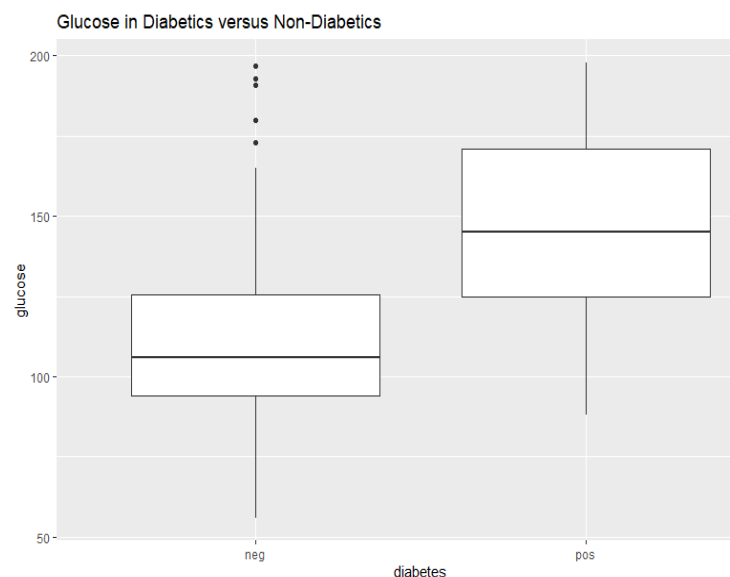
The mean and median glucose are higher in diabetics. This is likely due to type 2 diabetes being a condition where diet is often considered a prominent factor. [\[2\]](#)

Glucose levels in nondiabetics feature more outliers and a greater range but much of the data is in a relatively compact region of values. Meanwhile, in diabetics we see that glucose levels are spread almost evenly with no region of values standing out as the region most diabetics fit into.

Neither distribution is normally distributed, and their variances are unequal so further comparisons between the two will have to be considered in their approach. Overall, the distributions have little in common.



Above are the graphs for the plasma glucose concentration in nondiabetics and diabetics. Data for nondiabetics are shown in the top graph and diabetics are shown in the bottom.



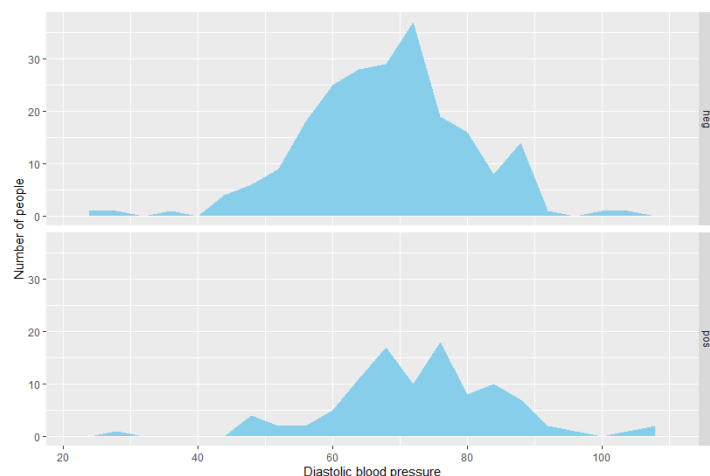
Blood Pressure in those with and Without Diabetes

Pressure	Diabetics	Nondiabetics
Mean	73.59259	69.04245
Standard Deviation	13.14053	11.89739
Median	74	70
Q1 and Q3	65.75, 82	62, 76
IQR	16.25	14

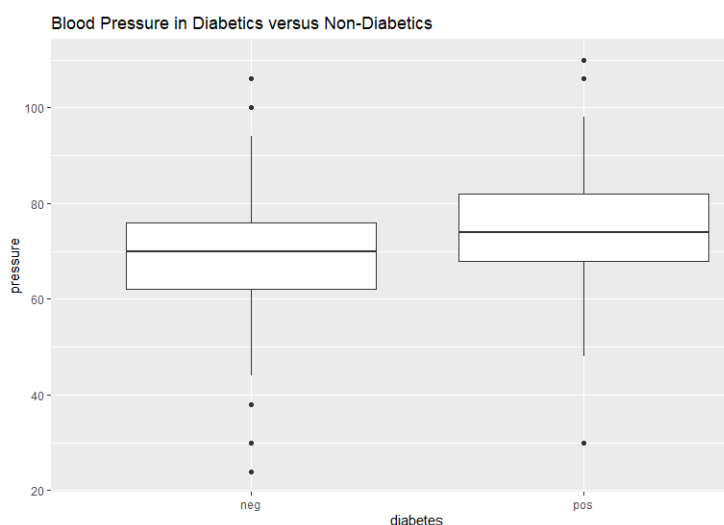
In the diabetic category, the mean and median blood pressure is higher. Perhaps this is because blood pressure and diabetes are both connected to stress hormones in the body.^[2]

Both groups include several outliers. A Shapiro-Wilk normality test returned that both distributions are normally distributed ($p=0.233$ for diabetics and $p=0.075$ for nondiabetics) and Levene's test for homogeneity of variance showed that the two distributions have equal variance. ($p=0.649$)

These distributions are broadly similar with the two having approximately equal variance and both being normally distributed. The only notable difference is that blood pressure is higher on average in diabetics. I did not conduct a hypothesis test to further explore the difference in means.



Above are the distributions for nondiabetics and diabetics, with the nondiabetics on top. It is visually apparent that the two distributions have a similar shape.



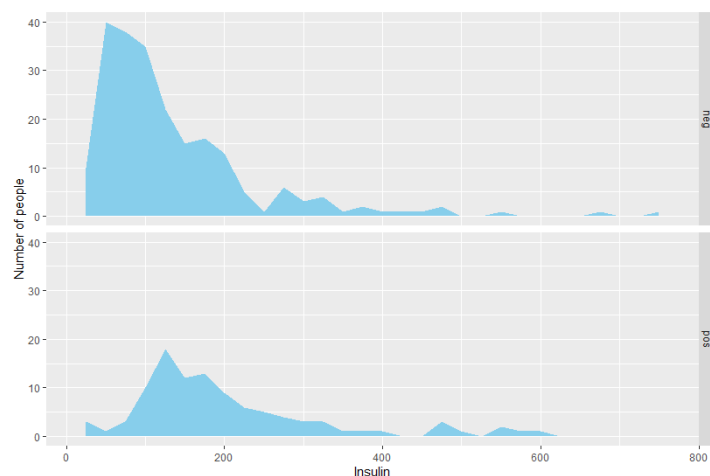
Insulin in those with and Without Diabetes

Insulin	Diabetics	Nondiabetics
Mean	205.5185	135.2453
Standard Deviation	138.0629	105.9256
Median	171.5	105.5
Q1 and Q3	121.5, 233.2	66, 166.5
IQR	111.7	100.5

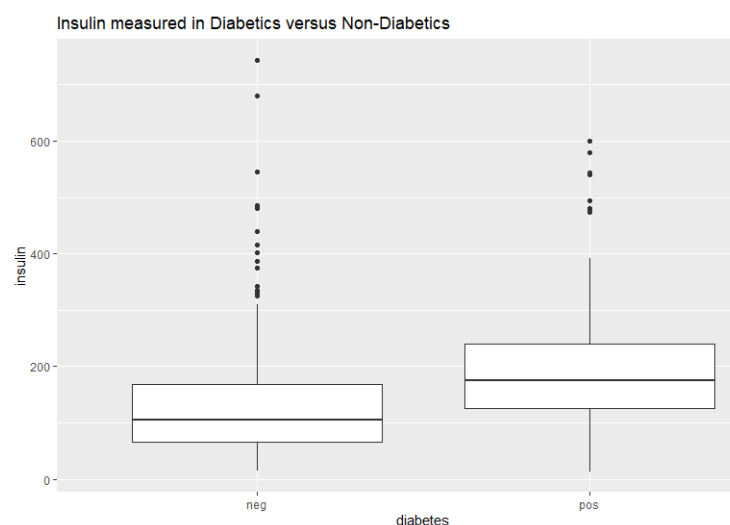
Insulin levels are higher in the diabetic group. If the prominent type of diabetes in the sample is type 2 then this makes sense as their body is producing its maximum amount without that being enough to regulate their blood sugar.

We see outliers in both groups, primarily outliers that are much larger than the rest of the data. Neither group is normally distributed but the two have similar positive skews. Levene's test reveals that the variances are approximately equal. ($p=0.184$)

The distributions in question certainly have many similarities such as positive skew and equal variances, but the values for diabetics seem noticeably higher. Due to the skewness, any further analysis of the differences between these two groups should use the median as the measure of centrality.



Above are the distributions for nondiabetics and diabetics, with the nondiabetics on top. Neither appear normally distributed and we observe a positive skew in both.



The Difference in Mean levels of Glucose in Diabetics and Nondiabetics

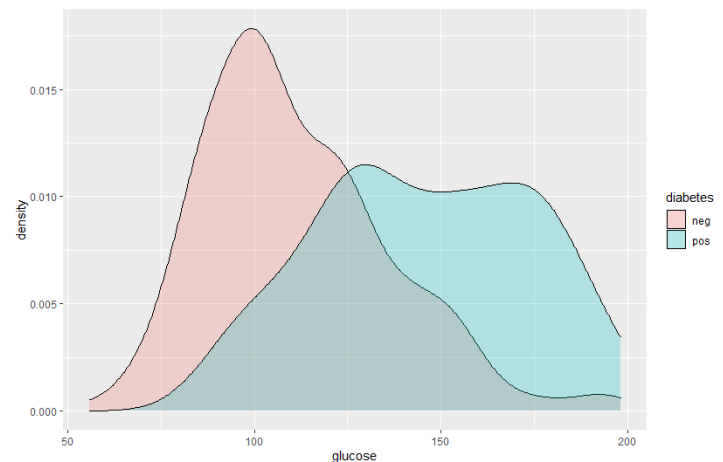
I initially sought to deduce which test for difference in mean of two groups would best work for this inquiry. A Shapiro-Wilk test applied to the distribution in both groups revealed that neither is normally distributed. ($p=0.0138$ for diabetics and $p=0.0002$ for nondiabetics)

This led me to believe a Mann-Whitney U Test or an Independent samples t-Test would work best. Levene's test for equality of variance revealed that the variances of the two groups were not equal. ($p=0.014$)

Under other circumstances I would use a Mann-Whitney U Test when it comes to the difference in these two groups because of their nonnormal distribution and inequality of variances. However, we are seeking to learn whether there is a statistically significant difference in the means and not the medians, so I used an Independent Samples t-Test.

I received a p-value of $2.2e^{-16}$ from the t-Test, meaning that the difference in means is significant. The 95% confidence interval for the difference is $[-40.78083, -27.74523]$.

I also carried out a Mann-Whitney U Test to satisfy my curiosity. It also returned a p-value of $2.2e^{-16}$ and so we are led to believe that there is a statistically significant difference in medians also.



Here we see the two distributions overlaid, with the red/pink colour being nondiabetics and the blue being diabetics. It is visually apparent that there is a sizeable gap between the means and medians of each group.

The Data from a Model Fitting Perspective

I sought to develop a model with the risk score as the dependent variable using the 8 other variables as independent variables. The correlation coefficients between each pair of variables were calculated, firstly for the Pearson correlation coefficient and then for Spearman's correlation. The Pearson coefficient correlations between risk and other variables were as follows,

Variable	Correlation Coefficient
Pregnant	0.158
Glucose	0.279
Pressure	0.834
Triceps	0.511
Insulin	0.180
Mass	0.743
Age	0.257
Diabetes	0.337

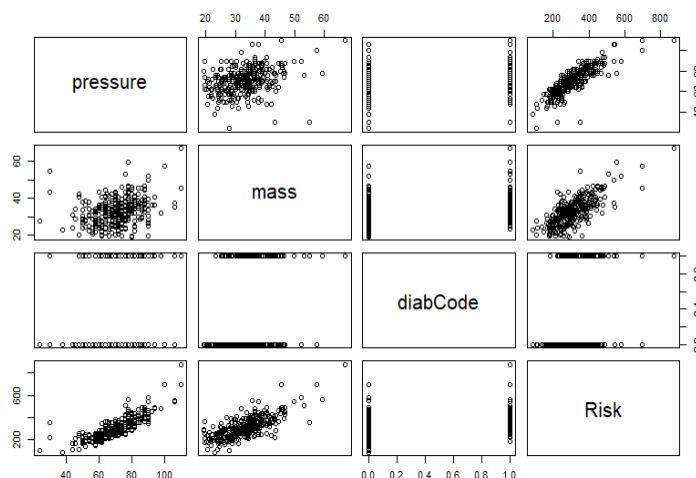
And for Spearman's correlation coefficient, we have,

Variable	Correlation Coefficient
Pregnant	0.123
Glucose	0.305
Pressure	0.855
Triceps	0.535
Insulin	0.230
Mass	0.711
Age	0.344
Diabetes	0.350

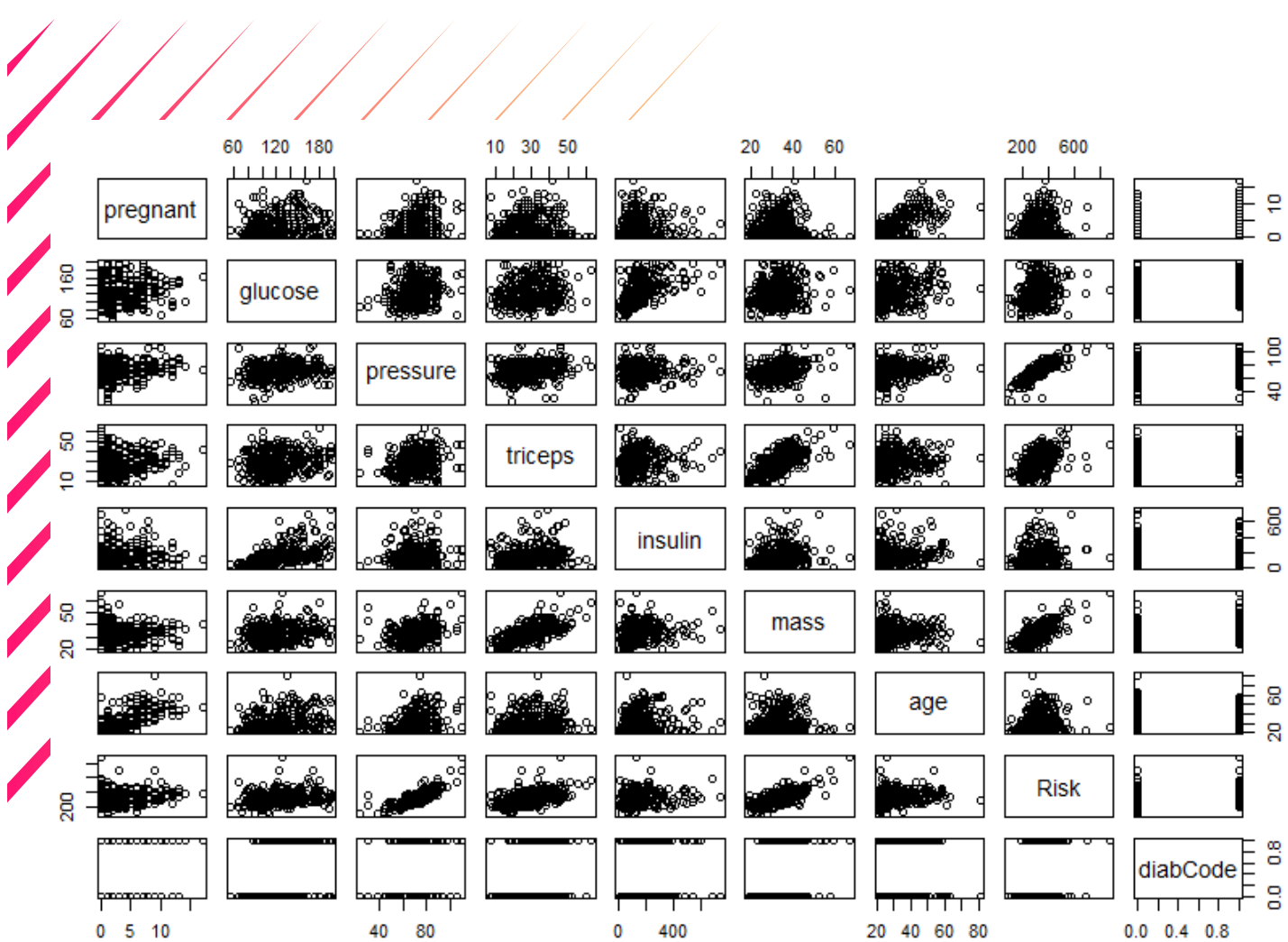
Both Pearson and Spearman returned similar values for each correlation

The scatterplots of each pair of variables are found on the next page. "diabCode" is simply a recoded version of the diabetes variable with 1=positive and 0=negative.

During the modelling process, I discovered that mass, pressure, and diabetes had the greatest effect on risk score. The scatterplots for those pairs were as follows,



The strong positive correlations for pressure and mass are visually obvious from these scatterplots.



The Data from a Model Fitting Perspective

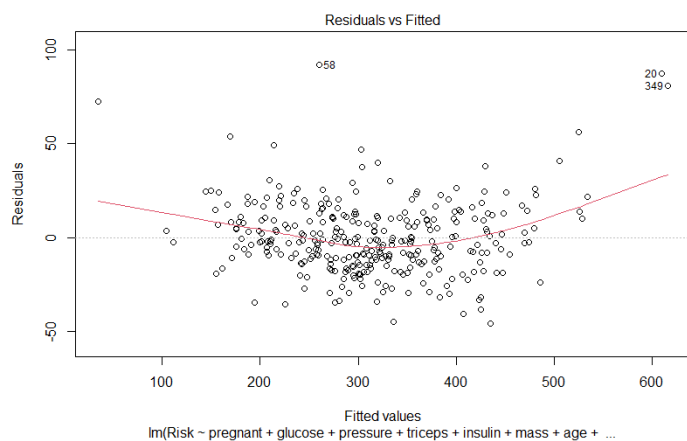
The first model I tried was simply letting risk score be linearly dependent on every other variable. i.e., my code was

```
model1 <- lm(Risk ~ pregnant + glucose + pressure +  
triceps + insulin + mass + age + diabCode, data = mydata)]
```

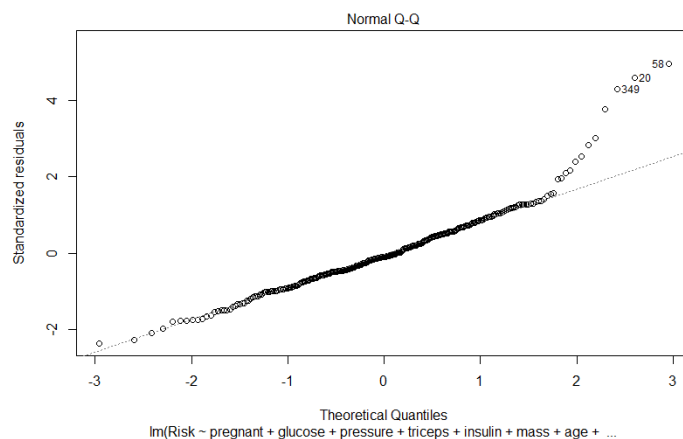
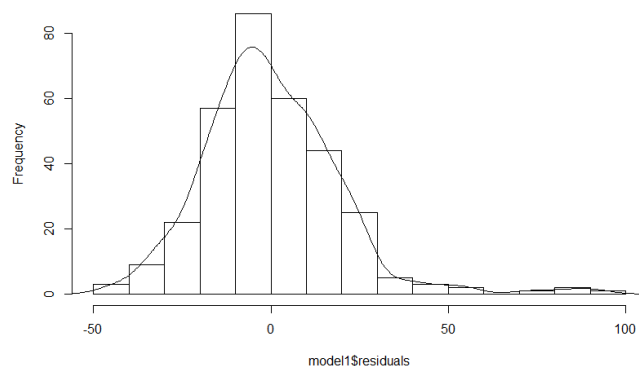
I did not have high hopes for this model, but I decided it would be a good starting point to improve upon.

The R-squared surprised me as even for this simple model, it was 0.9547. (or 0.9536 for the adjusted value) The model was statistically significant ($p=2.2e^{-16}$) and so this simple model proves itself as a good starting point.

However, the residuals were not only poorly fitted but they were also very nonnormally distributed. It was clear that this was not just the result of outliers but that the model was flawed in the importance it gives to its independent variables.



Above we observe the poor relationship between residuals and fitted values.



The above histogram and QQ-plot of the residuals shows their nonnormal distribution. The residuals are positively skewed and have many outliers on the upper end of their range.

Overall, this model is a good starting point but places too much importance on some variables and too little on others e.g., pressure and pregnant having the same importance when pressure and risk score have a much higher correlation than pregnant and risk score. There are also outliers that seem to be affecting the model that could be removed. I also have not addressed the relations between independent variables and their effect on the model, for example, triceps and mass.

The Data from a Model Fitting Perspective

After trying several approaches to improve the model, this is the model I eventually came to.

The first change is that I excluded any data that had a risk score either 1.5 times the IQR above or below the median risk score. I did this to exclude any women sampled who may have had extreme risk scores due to conditions which were either partially or entirely not accounted for in the data available to me.

I experimented with giving the variables different levels of importance and decided to remove several of the variables that seemed of little importance to the previous model. I retained pressure, mass, and diabetes. (in the form of diabCode) The most importance was placed on pressure, then mass, and finally diabCode. The code used to generate this model was

```
model3 <- lm(Risk ~ I(pressure^2) + I(mass^1.5) + diabCode, data = mydata.noOutliers)
```

There was not a notable improvement of the R-squared of this model. (0.9552 for this model as compared to 0.9547 for the previous model)

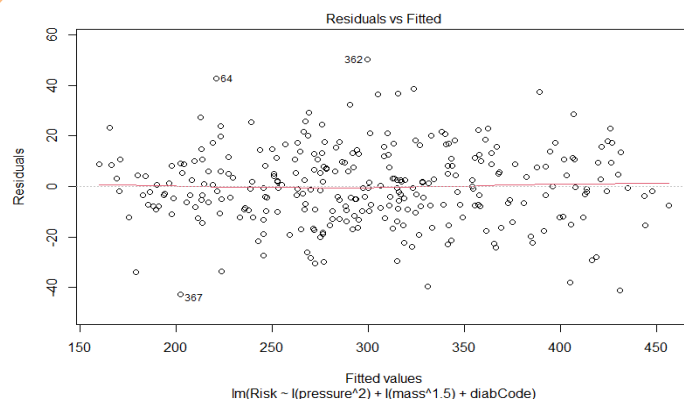
Below is the output for the model's summary.

```
Call:
lm(formula = Risk ~ I(pressure^2) + I(mass^1.5) + diabCode, data = mydata.nooutliers)

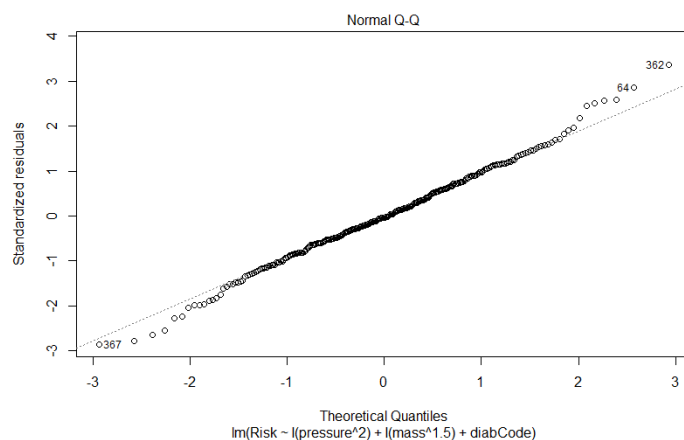
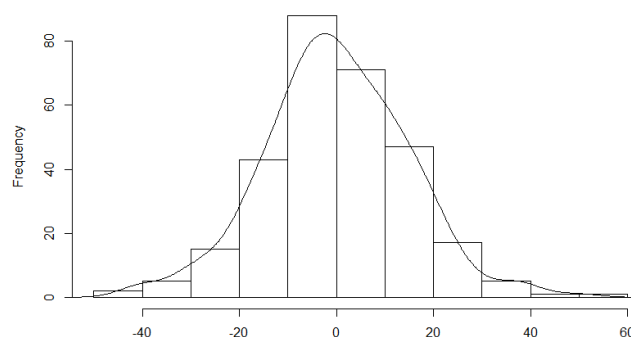
Residuals:
    Min       1Q   Median       3Q      Max
-42.711  -9.081   -0.585    9.679   50.318

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.097e+01  4.161e+00  -2.636  0.00885 **
I(pressure^2)  3.572e-02  6.163e-04  57.968  < 2e-16 ***
I(mass^1.5)    6.875e-01  1.702e-02  40.398  < 2e-16 ***
diabCode       1.413e+01  1.975e+00   7.156  6.79e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15 on 291 degrees of freedom
Multiple R-squared:  0.9552,    Adjusted R-squared:  0.9547
F-statistic: 2066 on 3 and 291 DF,  p-value: < 2.2e-16
```



The major improvement in this model can be seen above. The residuals are vastly decreased and better fitted in this model.



It is also clear that the residuals are more normally distributed in this model. I used a Shapiro-Wilk Test to verify that they are normally distributed. ($p=0.4696$) It is worth noting that transforming the dependent variable via a log transform did not improve the distribution of the residuals but was attempted.

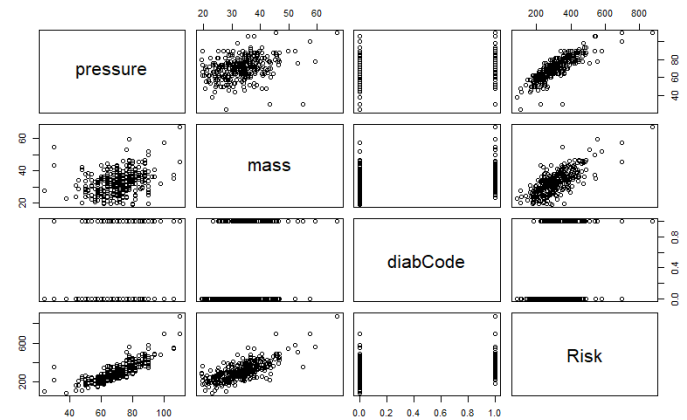
It is hopefully apparent why I choose to remove some variables from the model, for example, I excluded pregnant because its positive correlation with risk score is relatively small compared to variables such as pressure. My logic in removing triceps was that there is a strong positive correlation between mass and triceps ($r=0.68$ using Pearson) as well as both having a moderate/strong correlation with risk score. Both are similarly behaved factors of risk score and given the relationship between the two factors, it seemed that including both in the model would overestimate their importance. Upon investigating a model where importance was placed on both, it disimproved the spread and magnitude of the residuals.

The parameters of my model are based on a combination of the correlation coefficients and observing the change in the fit and spread of residuals as I went through a trial-and-error process with the parameters.

The strongest correlation between risk score and another variable is with pressure and that is why it is given the greatest amount of emphasis in my model.

BMI, or mass, had a weaker correlation than pressure but still significantly affected risk score. I also gave more emphasis to mass once triceps was removed from the model for reasons I have discussed previously.

Whether or not the woman was diabetic had enough correlation with risk score that I could not exclude it from the model, but it also did not exert as much importance as pressure or mass.



The variables included in the final model all have a positive correlation with the risk score.


We understand from these relations that the level of risk of a female Pima Indian's health is dependent on factors such as her BMI. This may be due to the health implications of being overweight such as more likelihood to have a stroke and the relationship between obesity and cancer.^[3]

Blood pressure has the strongest correlation with the risk score. High blood pressure also carries several health implications such as heart failure and kidney disease.^[4] Diabetes and high blood pressure are both also related to high levels of stress which might suggest that women in our sample with stressful lifestyles might be putting themselves at higher risk of their health failing them.

Diabetic women are more likely to have a higher risk score, and this is because diabetes has many complications associated with it from glaucoma to kidney disease. About 40% of diabetics develop nephropathy, the deterioration of proper functioning of the kidneys.^[5]



Citations

- [1] The NHS on BMI - <https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/>
 - [2] Large online community of diabetics - <https://www.diabetes.co.uk/causes-of-type2-diabetes.html>
 - [3] The CDC on the effects of being overweight/obese - <https://www.cdc.gov/healthyweight/effects/index.html>
 - [4] The American Heart Association on complications related to high blood pressure - <https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure>
 - [5] <https://www.diabetes.co.uk/diabetes-complications/diabetes-complications.html>
- 
-