Pima Indian Diabetes

Statement of Goals

Diabetes is a group of chronic diseases that leads to a high sugar level in the blood. It occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces [2]. Insulin is a hormone that regulates the blood sugar level in the body. Hyperglycemia, or raised blood sugar, is a common effect of uncontrolled diabetes. Over time, it leads to serious damage to many systems of the human body, especially the nerves and the blood vessels. With the `Pima Indian Diabetes` dataset, we want to explore, understand, and establish a relationship between the explanatory variables and diabetes mellitus in Pima Indian females. With this project work we want to answer the following:

- What are the most important factors/variables that can explain the diabetes mellitus in Pima Indian females?
- With the help of the explanatory variables, we wish to build a model that can predict the probability of a Pima Indian female being diabetic/non-diabetic.
- We also want to explore, whether individual variables are enough to build such a model, or do we need any interaction terms that can give more explanatory power to the model?

This report has been organized as follows:

```
i : Statement of Goals
1 : Historical perspective
2 : Description of Data
3 : Data Exploration and Modeling
4 : Making Modeling Choices
5 : A model with the top three variables(larger sample size)
6 : A model with all variables as predictors (larger sample size)
7 : A model with the overall most important predictor variables(reduced sample size)
8 : Conclusion, Limitations and Future Work
```

1. Historical perspective

The O'odham (Arizona), O'ob also Pima Bajo (Mexico) [4], or Pima in general, are descendants of the ancient Hohokam, who have inhabited the Sonoran desert and Sierra Madre regions for centuries[1] The Pimas of Arizona adapted to their desert homeland by directing water through an elaborate system of irrigation canals to support subsistence agriculture; they grew corn, beans, squash, and cotton. But due to the increased population of white settlers, these Pima Indians had to change their way of living. Their lifestyle changed and these changes included less physical labor and they faced scarcity of food. Their low-fat, high-carbohydrate diet changed to one that ultimately derived more than forty percent of its energy from fat[5]. This period coincides with the rise of diabetes in the people of this heritage. By 1970, the prevalence of type 2 diabetes was about forty percent among Pima Indians aged thirty-five and older. It currently affects about half of all Pima Indians over the age of thirty-five. To study these abnormal changes in diabetic people, many surveys were conducted and this data set is a part of one such survey.

Why should we care?: As per the current world scenario, diabetes is one of those silent diseases which kills thousands of people every year. According to the World Health Organization, around 1.6 million people worldwide died due to diabetes in 2016 [1]. The number of people with diabetes has risen from `108` million in 1980 to `422` million in 2014 [2]. As of 2019, there were `463` million diabetic people on this planet and by 2045, the projection shows that this number may rise to some `629` million people globally [1]. This makes it important that we explore a suitable dataset like, Pima Indian diabetes and try to understand what leads to a sudden increase in the number of diabetic people in a community that wasn't historically prone to this

disease.

2. Description of Data

This dataset is taken from the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of several independent medical explanatory variables and one target variable, `Outcome`. The data was collected through a number of surveys done across the Pima Indian community in Mexico and the United States. It was aimed to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. It contains information of 768 women over the age of 21, belonging to the Pima Indian heritage. The density plot of each variable is shown in figure 1 and a brief description of each of these variables is as follows:

Age:- It is the count of a person's age in years. The distribution of `age` is right-skewed, with very less number of people above `60` years. Taking a logarithmic transformation could not improve the distribution, hence, we left this distribution as it is. The minimum age is `21` years, and the maximum age is `81` years with a mean value of `33` years.

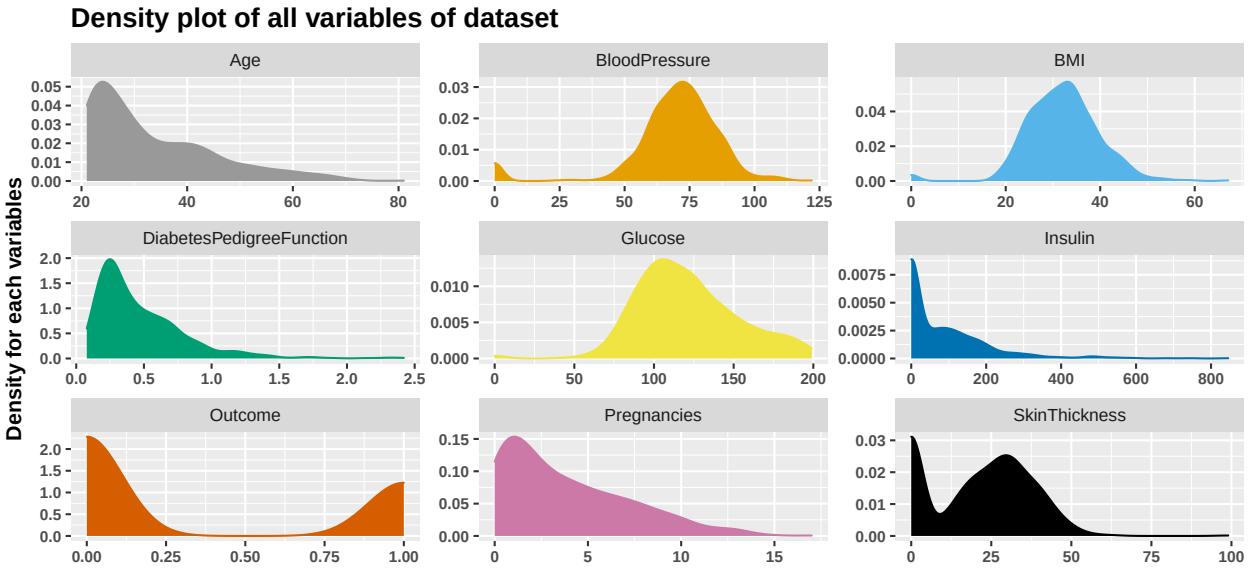


Figure 1: Density plot of all variables as given in the dataset.

Blood Pressure:- It is the `diastolic` blood pressure (mm Hg) reading of females. `Systolic blood pressure`, i.e., the first number in the measurement of blood pressure, indicates the amount of pressure the blood exerts against the wall of a person's arteries during heartbeats. `Diastolic blood pressure`, i.e., the second number in the measurement of blood pressure, indicates the amount of pressure the blood exerts against the wall of a person's arteries when the heart is resting between beats. The maximum reading is `122` mm Hg and the minimum reading is `0`, which are the missing values. There are thirty-five such values. The distribution of this variable is suitable for our work.

Body Mass Index (BMI):- It is the BMI reading of females. It is the ratio of weight (kg) to the square of the height (m). It has eleven missing values which are indicated by `0`. The maximum BMI is `67.10` and the mean value is `32.` This variable is positively right-skewed. So, we have taken a log transformation to improve the distribution.

Diabetes Pedigree Function:- It is a likelihood score of diabetes, based on a person's family history. It provides some insight into the diabetes mellitus history in relatives and the genetic relationship of those relatives with the patient. This measure gives an idea of the hereditary risk of becoming diabetic for an individual. We have taken a log transformation of this variable as it was right-skewed.

Glucose level:- It contains the reading of plasma glucose concentration- a 2-hour oral glucose tolerance test. The mean value of this variable is `120`, the maximum value is `199` and the minimum value is `0` (5 such values). The zeros are the missing values, i.e. reading was not taken for that particular person. It is measured in mg/dL(milligrams per decilitre).

Insulin level:- It is the numerical measurement of the outcome of a 2-hour serum insulin test, measured in μ U/ml. The numbers indicate the insulin level of a person after two hours of a meal. The minimum value is `0` and the maximum value is `846` with a mean of `79.`
Outcome:- It is the response variable, which indicates if a particular person has diabetes or not. It is coded as a binary outcome where `0` means, the person doesn't have diabetes and `1`, means the person has diabetes. `268` of `768` `Outcome` variables are labeled as `1`, and the others are labeled `0`.
Number of Pregnancies:- It is a number signifying the number of times the female was pregnant. The lowest value for this variable is `0` and the highest value is `17` and it has the median value of 3.
Skin Thickness:- It is a measurement of triceps skinfold thickness(mm). There are `227` missing values for this variable, indicated by `0`. The maximum value corresponding to this variable is `99` mm, and a mean value of `21` mm.
Data Sampling method: The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), conducted a cross-sectional study to identify the effects of traditional and western environments on prevalence of type 2 diabetes mellitus and obesity in Pima Indian community in Mexico and the United States. This study was conducted twice, once in 1995 and then in 2010.[6] The method used in 1995 and 2010 was the same, barring the change in the sample size. Measurements of weight, height, body fat (bioimpedance), blood pressure, plasma levels of glucose, cholesterol, and HbA1c were obtained in women (36 +/- 13 years of age) and men (48 +/- 14 years of age) [7]. We focus only on the data of women with more than 21 years of age, which is a subset of larger survey data.

3. Data Exploration and Modeling

We start with the broader choice of which kind of model can be used on this data. We explore the dot plot of each variable with the outcome variables and subsequently try to observe patterns by fitting a loess curve, the loess curve generates a neighborhood average to draw a line. The plot is shown in figure 2. This figure

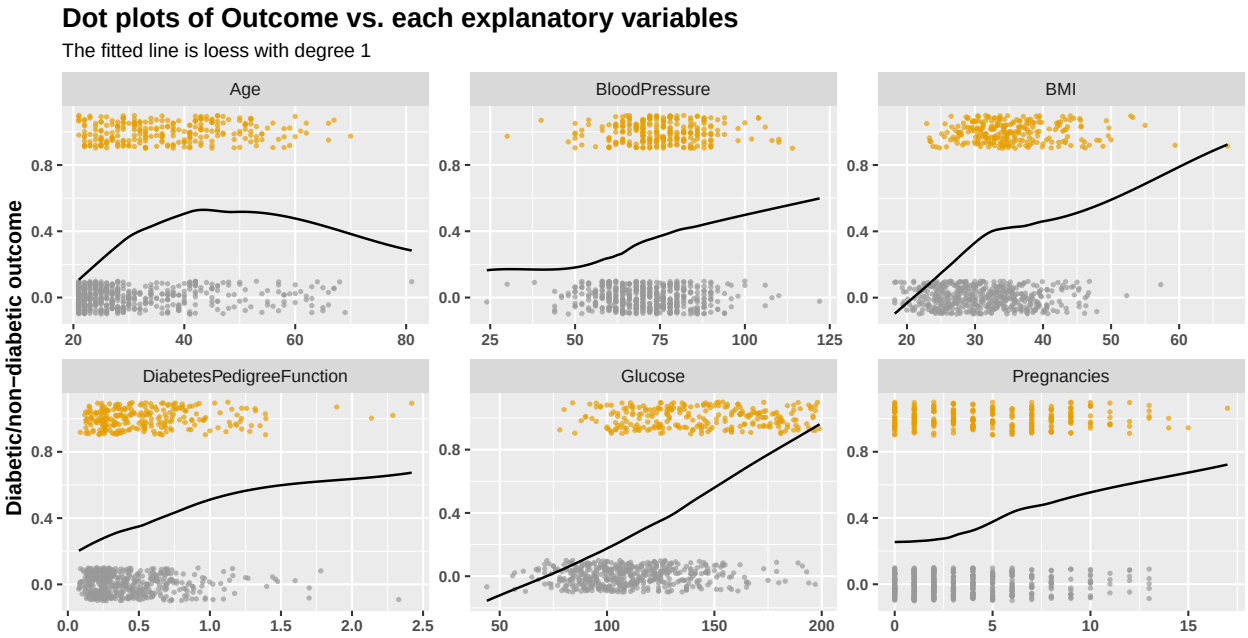


Figure 2: Fitting loess model to observe the behaviour of outcome with each explanatory variables

shows each variable plotted against the response variable, the black line is the loess curve. For the `Age` variable, the loess model shows an increasing pattern with age, but for higher values of age, we do not have enough samples to observe a consistent pattern. Also, it seems that the line is largely bent due to one outlier in the data. For the `Blood pressure` variable, the pattern is almost straight and increasing after an initial slow rate of increase, mostly due to fewer samples in the region. For the `BMI` variable, we can see that the loess model shows an increasing pattern, but we have fewer samples for BMI values greater than 45. We will have to be careful while modeling. Also, the loess line has been affected sharply by the few outliers

beyond the BMI value of 50. For the `Diabetes pedigree function` variable, the pattern observed can be well modeled by logistic regression. There are some outliers for this variable, so while observing the model, we should be aware of the outliers and try not to comment on any such region which has fewer samples. `Glucose` shows a perfect pattern that can be captured well by logistic regression. It seems from the plot that this would be the most prominent variable in decision making. `Number of pregnancies` is also showing an increasing pattern well suited for logistic regression. Overall, we have enough evidence to model this data by logistic regression.

4. Making Model Choices

After exploring each variable, we came up with the following modeling decisions:

- Variables like `Blood Pressure`, `BMI`, and `Glucose` have a few missing values. So, we decided to drop these missing values. A total of 44 such samples were dropped.
- Variables like `BMI` and `Diabetes Pedigree function` are positive and right-skewed, and log transform improves their distribution, so we take log transformation for these two variables.
- The variable `Age` is also positive and right-skewed, but taking log transformation doesn't help much with the distribution. So, we forgo the transformation.
- The variable `Pregnancies` is also highly right-skewed. It has a lot of meaningful zero values, which we can't replace with any other meaningful number. This makes it difficult to transform on a log scale, so, we will use this variable as it is.
- Two of these variables, `Insulin` and `Skin thickness` have a lot of zeros in them, which are the missing values and not relevant. So, we decided to make models with two different sample sizes. First, we considered only six of the eight explanatory variables (dropping Insulin and Skin Thickness) with a sample size of 724, and then, in the second model, we considered Insulin and skin thickness as well, but with a reduced sample size of 392. The sample size is reduced as we remove all missing values in Insulin and Skin Thickness.

Other models that we tried: Apart from the above-mentioned modeling choices, we also tried many different options, like, we converted the Blood pressure level into three categories as we have a clear classification for systolic blood pressure levels [3]. But this did not improve our model. Then, we tried to segment age with a 10-year gap between the groups, just to explore, if these changes somehow affect the model performance or if there is some systematic pattern between these groups, but we did not find anything interesting in that model. We also tried GAM (Generalised Additive Model) with as well as without interaction terms, but no significant improvement was observed in any of these models. So, we decided to stick to a simple model with better interpretability.

We start our study by observing the Pearson's correlation coefficient among all six variables. Table 1 shows the correlation values for these variables. We observe a significant correlation between a person being diabetic

Glucose	BloodPressure	log.BMI	log.DFP	Age	Outcome	
0.135	0.21	0.035	-0.0247	**0.557**	0.224	Pregnancies
	0.223	0.231	0.112	0.264	**0.488**	Glucose
		0.277	-0.002	**0.325**	0.167	BloodPressure
			0.142	0.037	**0.309**	log.BMI
				0.0216	0.188	log.DFP
					0.246	Age

Table 1: Correlation between different variables (DFP: Diabetes pedigree function)

and the `glucose` level, which seems to follow the general medical understanding of this disease. A less technical definition of a person being diabetic is a condition with consistently high glucose level in blood. It should have a significant correlation value with the `outcome` variable, as is the case here. The log-transformed `body mass index` is also significantly correlated with the `outcome` variable, which is also following the general understanding to some extent. The variables `Age` and `Pregnancies` have a significant value of correlation,

which is not counterintuitive at all, in general, older females have more number of pregnancies. Other variables are also showing significant correlation, like `Blood pressure` with `Age` and `Blood pressure` with `Log BMI`. A person with a high BMI tends to have high blood pressure measurement [8] and also, blood pressure measurement tends to increase as a person's age increases [8,9]. All these correlation values are following our intuitive understanding of this disease.

5. A model with the top three variables(larger sample size):

With the above understanding of the data, we first picked those variables which have a high correlation with the `outcome` variable. Starting with a one-variable model, `Glucose`, we tried to find out the best predictors or best combination of predictors for this model. With Glucose as the only predictor, the model gave an accuracy of around 75 percent, but it was just a base model. So, we further tried a combination of variables to see if any of those combinations boosts the performance of the model. A combination of three variables, without any interaction between them, Glucose, Pregnancies, and BMI.log gave an accuracy of around 77.5 percent, on the same data. This suggests that these three variables are the leading predictors of diabetes mellitus on this data. This model explained about 31.99 percent of the variance in the data. The model equation is:

```
Diabetes.logit.1 = glm( Outcome ~ Glucose + Pregnancies + BMI.log , family = "binomial",
                  data = diabetes.log[c("Pregnancies", "Glucose", "BMI.log", "Outcome" )])
```

Prediction on new data: To better understand the model, we observed each variable on grid data. Glucose is our main predictor variable, so we made a dense grid in glucose level, varying from `75` to `200` with step `1.` We have taken a few values of Pregnancies, like 0, 3, 6, and 9, and settled for a mean value of BMI.log, `1.502.` We predicted the outcome on this new grid and collected the predicted probability values, i.e. response, and plotted it as shown in figure 3. The plot shows the change in the probability of being diabetic for a Pima

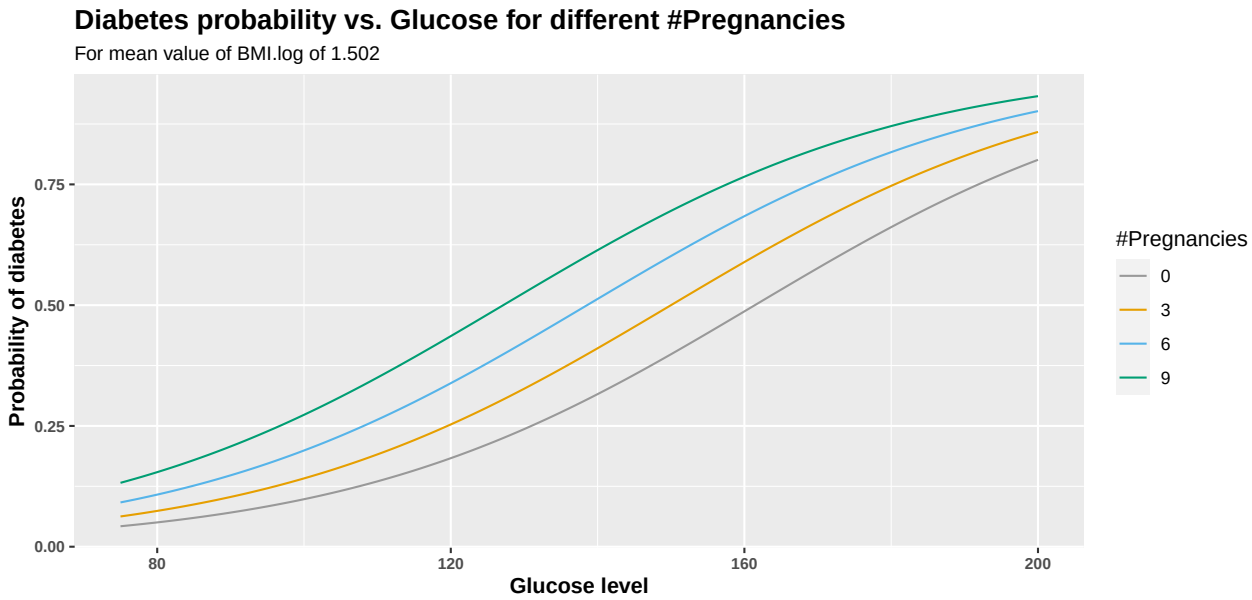


Figure 3: Observing the prediction probability on grid for different values of pregnancies

Indian female with different values of pregnancies and with the mean value of BMI.log set to `1.502.` The lines in the plot show a significant change in probabilities for women with a higher number of pregnancies. A female with no pregnancy and a log.BMI around `1.502`, is at a much lower risk of being diabetic for an even higher level of glucose (close to the value of `160`); but Pima Indian females with pregnancies between `6` and `9` are more prone to be diabetic with a glucose level of around `130` and `140.` This shows a significant effect of the number of pregnancies on the diabetes mellitus among Pima Indian females.

We further try to understand the pattern of probabilities of being diabetic for different values of the log.BMI with a change in Glucose level. This plot is shown in figure 4. The plot shows the prediction probability of

being diabetic for different values of the log of BMI with the change in Glucose level; when the number of pregnancy is fixed at its median value of `3`. For this plot, we have restricted the values in the grid only to the area where we have more data points. This has lead to the removal of extreme values from both ends of the log.BMI values. The lines corresponding to each log.BMI values are quite spread out from one another, which shows a significant variation in probabilities with different slope values, like, for lower values of log BMI, the increase in probability is significantly slower, but for log BMI values of `1.5` (which is equal to `31` on the original scale of BMI) the change in probability is significantly faster, the plot crosses the probability value of `0.5` around `150` glucose level. This value is much smaller for a log BMI value of `1.6` (equal to `39.8` on the original scale) at around glucose level `130`. This model significantly captures the different patterns of how a change in glucose level impacts different BMI values for diabetes mellitus.

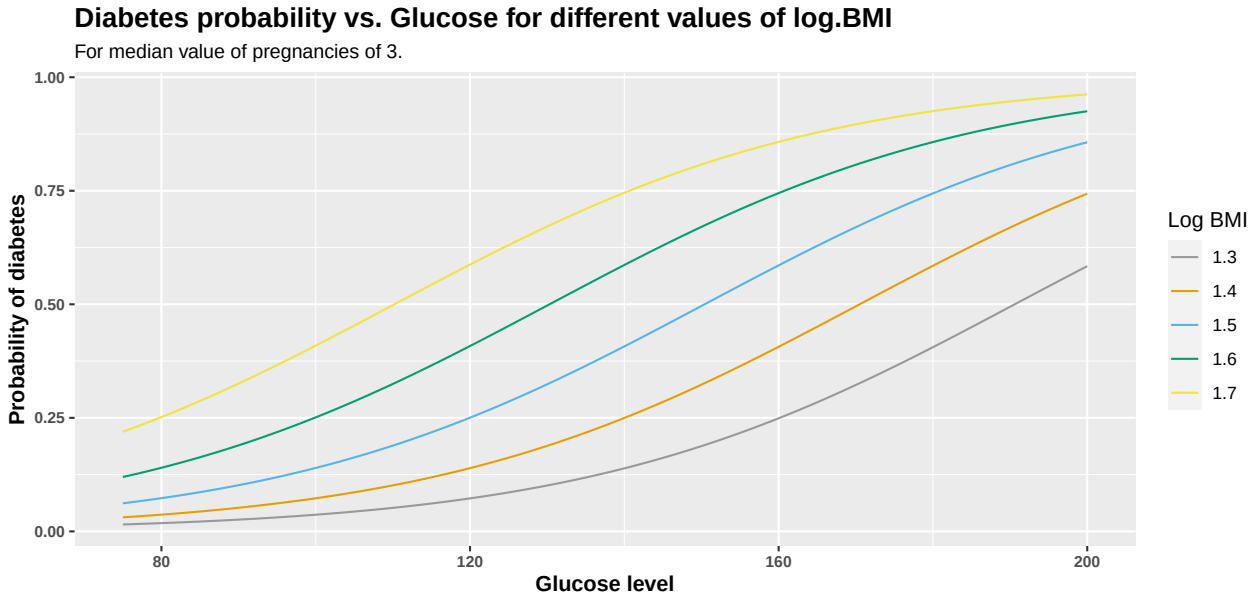


Figure 4: Observing the prediction probability on grid for different values of log.BMI

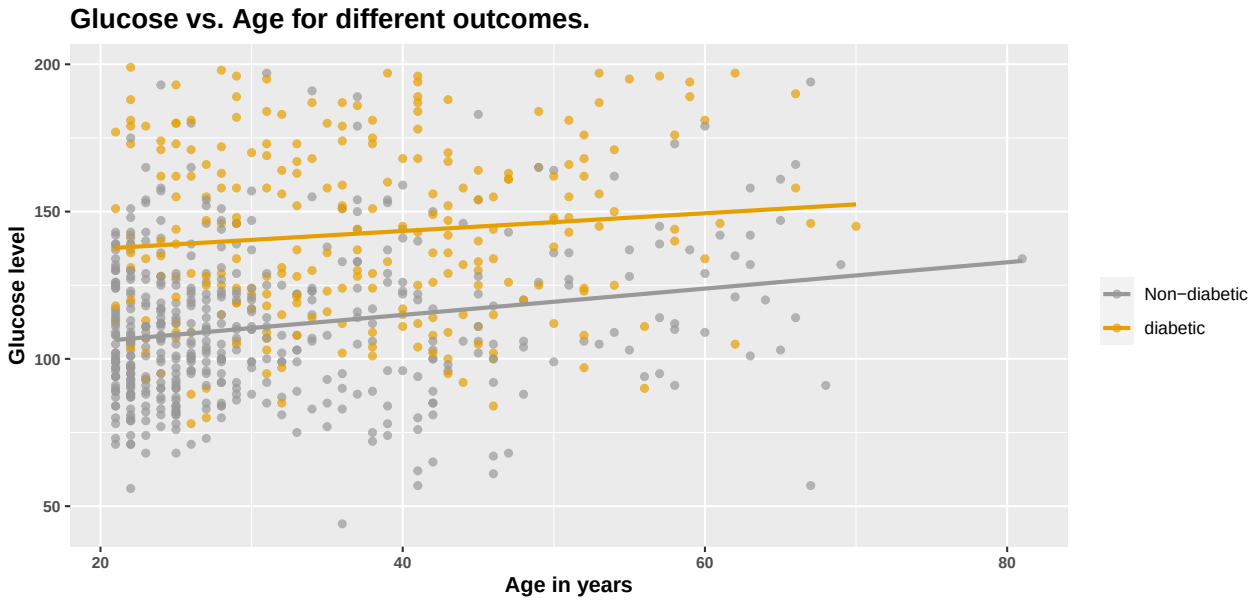


figure 5: Observing the interaction between Age and Glucose level with different outcomes.

The three-variable model helped us in developing an understanding of the difference in the probability of being diabetic with different values of pregnancies and different values of the log.BMI with a change in the glucose level in Pima Indian females. Now, we explore all six variables as predictors on a large dataset (sample

size 724) and try to understand the behavior of other variables as well.

6. A model with all variables as predictors (larger sample size):

We wanted to explore the effect of including all other variables as predictors in terms of model accuracy and the variance explained. We also wanted to explore, how the interaction between any of the two terms affects the model. We tried all permutations and combinations of interactions, starting with those variables which showed high correlation values with each other. But none of the interaction terms significantly improved the performance of the model or the variance explained, so we removed all interaction terms from the model except `Age:Glucose` as it positively affected the model. Figure 5 shows `glucose` vs. `age` plot with the color-coded outcome variable. The plot shows that diabetic females are more dominant in the high glucose level. Another important point to notice from this plot is that we don't have much data for higher values of age, so we should be careful about this while fitting a model. This model shows the accuracy of 78.5 percent on the same data and explains about 33.88 percent of the variance, which is a slight improvement upon the three-variable model. We get almost 1 percentage point improvement in accuracy over the three-variable model. The model equation is written below:

```
diabetes.logit.2 = glm(Outcome ~ Pregnancies + Glucose + BloodPressure + BMI.log +
                       DiabetesPedigreeFunction.log + Age + Age:Glucose,
                       family = "binomial", data = diabetes.log)
```

Prediction on new data: To develop a critical understanding of the model, we observed the model on grid data. We have already observed a few variables, like log-transformed BMI, the number of pregnancies with glucose levels in our first model. So, we do not show those plots again. Here, we only show observations on the other three variables, Blood pressure, Diabetes Pedigree function, and Age. With glucose as our main predictor, we make a dense grid in glucose levels, with blood pressure levels as 60, 75, 90, 105 and 120 mm Hg to observe the pattern with blood pressure as one of the factors. We keep all other variables fixed at their mean/median values. This plot is shown in figure 6. It shows the probability of being diabetic with a change

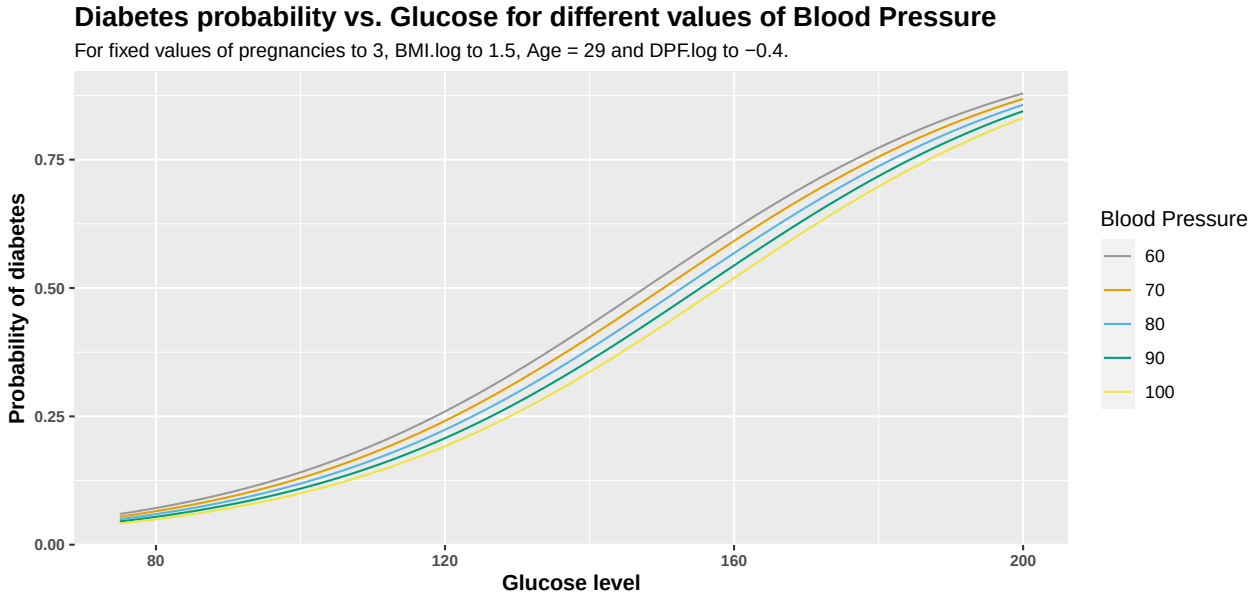


figure 6: Observing the prediction probability on grid for different values of blood pressure

in glucose levels for different blood pressure values. Corresponding to these two variables, we have more data points for blood pressure levels between 60 and 100, so we restrict our attention in that area only. The lines corresponding to the blood pressure values do not show a significant variation from each other. For a given value of glucose, we can see the lines shift very slightly along the x-axis. This indicates that the change in blood pressure level doesn't highly or directly affect the onset of diabetes in Pima Indian females.

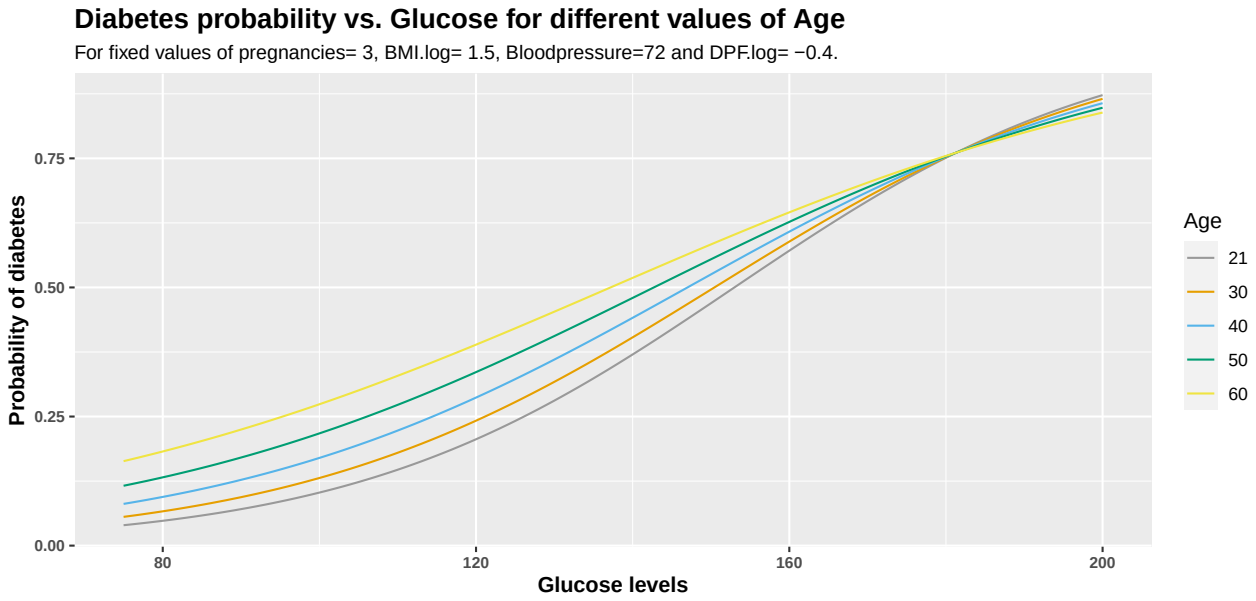


figure 7: Observing the prediction probability on grid for different values of Age.

Next, we try to understand how the relationship between Glucose and Age varies the diabetes mellitus prediction probability. This is shown in figure 7. We don't have enough data points beyond 60 for age, so we restrict our attention between 21 to 60 years of age. We chose five values of age, 21, 30, 40, 50, and 60 and make a grid dense in glucose level, keeping all other variables at their mean/median values. The lines corresponding to age 30, 40, and 50 show a rapid increase in probability with glucose level. This indicates that the chances of being diabetic increases sharply between 30 to 50 age group. This was also suggested in the original study. For higher aged females (with age 50 and 60), the probability of being diabetic is slightly less as compared to others, but it is still significantly high. This observation suggests that Pima Indian females with high glucose levels, belonging to any age group will have a high probability of being diabetic.

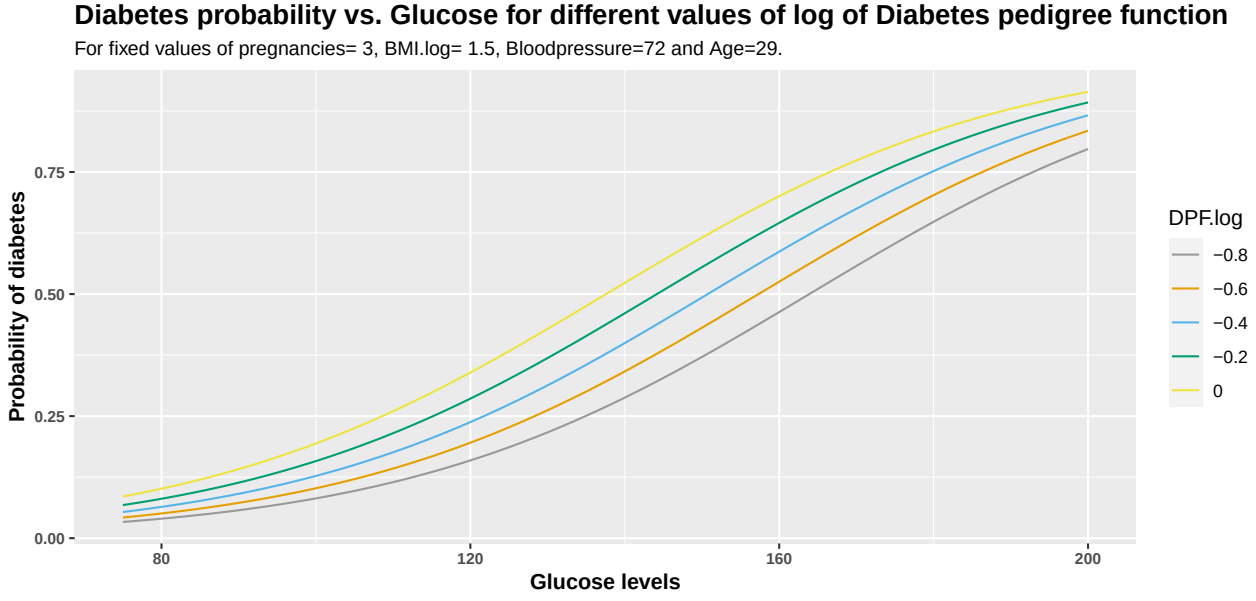


figure 8: Observing the prediction probability on grid for different values of log of Diabetes Pedigree function.

The next two variables that we want to observe are Glucose and log of Diabetes pedigree function. This is shown in figure 8. We restrict our attention to only those areas where we have sufficient data to study the pattern. With a dense grid in glucose level and with few values of diabetes pedigree function, we see that each line corresponding to the log of the pedigree function traces a slightly different path, shifted along the

x-axis. The plot shows that the chances of getting diabetes increases slightly with the increase in the value of diabetes pedigree function in Pima Indian females. This indicates that there might be some heretical factors involved, but this particular factor is not so strong in our data set, which we also observed in the correlation table 1.

While making the initial modeling decisions, we dropped Insulin and skin thickness from our dataset due to several meaningless zeros (around 45% of samples) present in them. But we also wanted to study the effect of including `Insulin` and `Skin thickness` explanatory variables and study their impact on the explanatory power of the model.

7. A model with the overall most important predictor variables(reduced sample size).

First, we prepare a clean data by dropping all meaningless zeros. It reduces our sample size to `392.` Apart from adding two variables, we keep all other modeling choices the same. Again, we start with the Pearson correlation coefficient table for the variables. It is shown in table 2.

Glucose	BloodPressure	log.BMI	log.DPF	Age	Insulin	SkinThickness	Outcome	
0.198	0.213	0.0019	0.0234	**0.68**	0.079	0.0932	0.257	Pregnancies
	0.21	0.221	0.111	**0.344**	**0.581**	0.199	**0.516**	Glucose
		0.293	-0.0274	0.3	0.0985	0.233	0.193	BloodPressure
			0.123	0.088	0.239	**0.672**	**0.281**	log.BMI
				0.091	0.101	0.115	0.206	log.DPF
					0.217	0.168	**0.351**	Age
						0.182	**0.301**	Insulin
							0.256	SkinThickness

Table 2: Correlation between different variables (DFP: Diabetes pedigree function)

With 392 samples, most of the correlation values have slightly increased. `Glucose` shows the highest correlation value of `0.516` with the `outcome` variable, which is followed by the correlation between `Age` (`0.351`) and the outcome and the correlation between `Insulin` (`0.301`) and the outcome. Some other correlation values are also high, e.g. the correlation between `Age` and the number of `pregnancies`, but this doesn't affect our model. Another high correlation is between `Skin thickness` and `log.BMI`, which seems like a natural relationship and doesn't impact our model. This time, we wanted to focus on the best predictors only, though we tried creating a model with all eight variables. There wasn't any improvement in the model in terms of prediction accuracy or the variance explained. This model incorporates the four best variables that would act as predictor variables for this dataset. The model is represented below:

```
diabetes.logit.all1 = glm(Outcome ~ Glucose + Insulin + Pregnancies + BMI.log ,
                          family = "binomial",
                          data = diabetes.all)
```

In terms of coefficients, we can write the model equation as

$$logit(P(Outcome|Glucose, Insulin, Pregnancies, log.BMI)$$

$$= -16.73 + 0.039 * Glucose + 0.147 * Pregnancies - 0.006 * Insulin + 6.595 * Log.BMI$$

Between the two added variables, Insulin and Skin thickness, insulin seems to play an important role in improving the model. But skin thickness didn't do as much. A probable explanation for this could be, the skin thickness has a high correlation with BMI.log, and keeping both of them together didn't improve the model as such. This model gives 79 percent accuracy on the same data, which is highest among all the models. It captures almost 33.5 percent of the variance in data with only four variables. This certainly makes these variables the most important variables for predicting the probability of diabetes mellitus for a Pima Indian female.

Prediction on new data: For this model, we only show the variation for Insulin as we have already explored all other variables in the earlier models. Skin thickness didn't turn out to be important at all, so we will ignore it. We do not have high values of insulin level corresponding to the lower values of glucose levels, so, we had to trim the grid to observe the data in that range. The corresponding plot is shown in figure 9. The plot highlights the behavior for different levels of insulin. The abrupt end of lines is due to the trimming of values. The change in the probability of having diabetes is not significantly big for a higher level of insulin, but for the lower levels of insulin, the corresponding line has the largest slope, which is analogous to the condition when the body fails to produce enough insulin as the cause of diabetes. This shows a direct relationship between insulin and glucose and how they relate to diabetes mellitus.

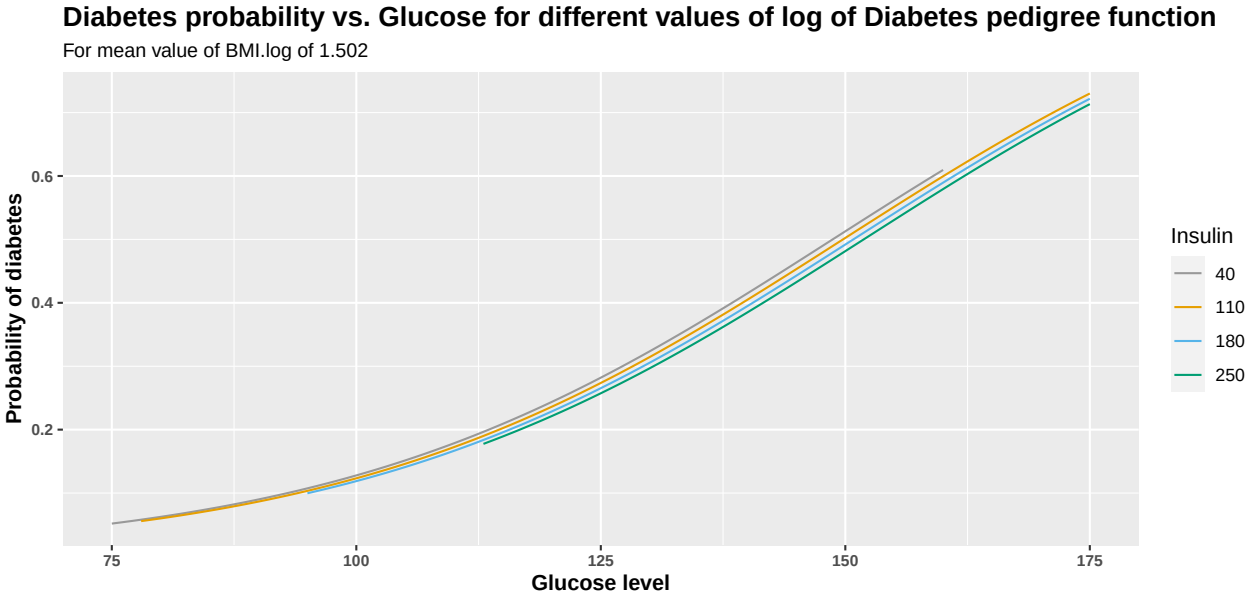


figure 9: Observing the prediction probability on grid for different values of Insulin.

8. Conclusion, Limitation and Future Work

Conclusion: Our extensive exploration has left us with some interesting insights about diabetes and its relationship with Pima Indian women. Our study suggests that Glucose is the main predictor variable. Amongst the other explanatory variables, we found BMI(log), number of pregnancies, and Insulin seemed to improve the fit of the model. We also explored all the possibilities of interaction between variables, but in the end, we chose a more interpretable logistic regression with no interaction terms for our best models. With our top three variables as predictors, i.e., Glucose, number of Pregnancies, and log of BMI, we could achieve the accuracy of about 77.5% and with the top four variables, with added Insulin variable, we could achieve the accuracy of about 79 percent. Studying each variable on new grid data has given more insights. The probability of diabetes varies significantly with the number of pregnancies in Pima Indian women, women with a higher number of pregnancies are at higher risk. Similar is the case with higher BMI values, the high-value BMI curve shows a rapid increase in probability with an increase in glucose level. But this was not the case with blood pressure values, for high as well as low values, the plots were not much different. For the age of 30 to 50, the slight change in glucose level rapidly increases the probability of being diabetic. At a low level of insulin, the probability changed quickly but that was not the same for a high level of insulin. Diabetes pedigree function didn't show a significant impact on probability values, this follows the original theory of increase in diabetes cases due to change in lifestyle and not due to some heretical influence. We believe that the model with Glucose, number of pregnancies, log-transformed BMI along with Insulin, is the best model amongst all and these variables are the best predictor variables to predict diabetes mellitus in Pima Indian females.

Limitations: This dataset also had some short-comings, especially with the readings of Insulin. With insulin as one of the main predictors, the lack of data for an important variable impeded our study to effectively conclude that insulin levels are strong predictors of a person being diabetic or not as suggested in [10]. As

mentioned in the introduction, the drastic rise in the number of diabetic people amongst the Pima Indians could be attributed to the changes in their lifestyle. Consequently, due to less physical labor, their usage of glucose decreased and affected insulin levels negatively. Since we did not have enough insulin data, we could not explore the effect of insulin levels that deeply. Apart from that, there are other factors too, like, the amount of physical activity performed by each subject per day. This type of data could contribute to the effectiveness of this study by indicating what brought about the sudden change in this community. Somewhere, we also felt that this study would have been more meaningful on a much larger dataset.

Future Work: This dataset captures some of the factors leading to diabetes very well. This study could be used as a baseline model to conduct a more in-depth study of this disease in the Pima Indian community. We don't need to restrict our study only to females as well. As a part of the future work, collecting more samples, especially with our four main predictor variables can give new insights. Since the Pima Indian community went through a lifestyle change during the same time when there was a rise in the number of diabetic patients in this community, some more information about a person's average physical activity per day, food habits, nutrition details, could reveal more details about what led to the rise in diabetes in the community. On the other hand, we could also perform a randomized controlled experiment with a group of Pima Indian people and try to establish a causal relation more firmly. One another thing to do would be, observing a bunch of people over time and then do a time-series analysis to explore different trends in the glucose levels, obesity, nutrition that leads to diabetes mellitus in the Pima Indian community.

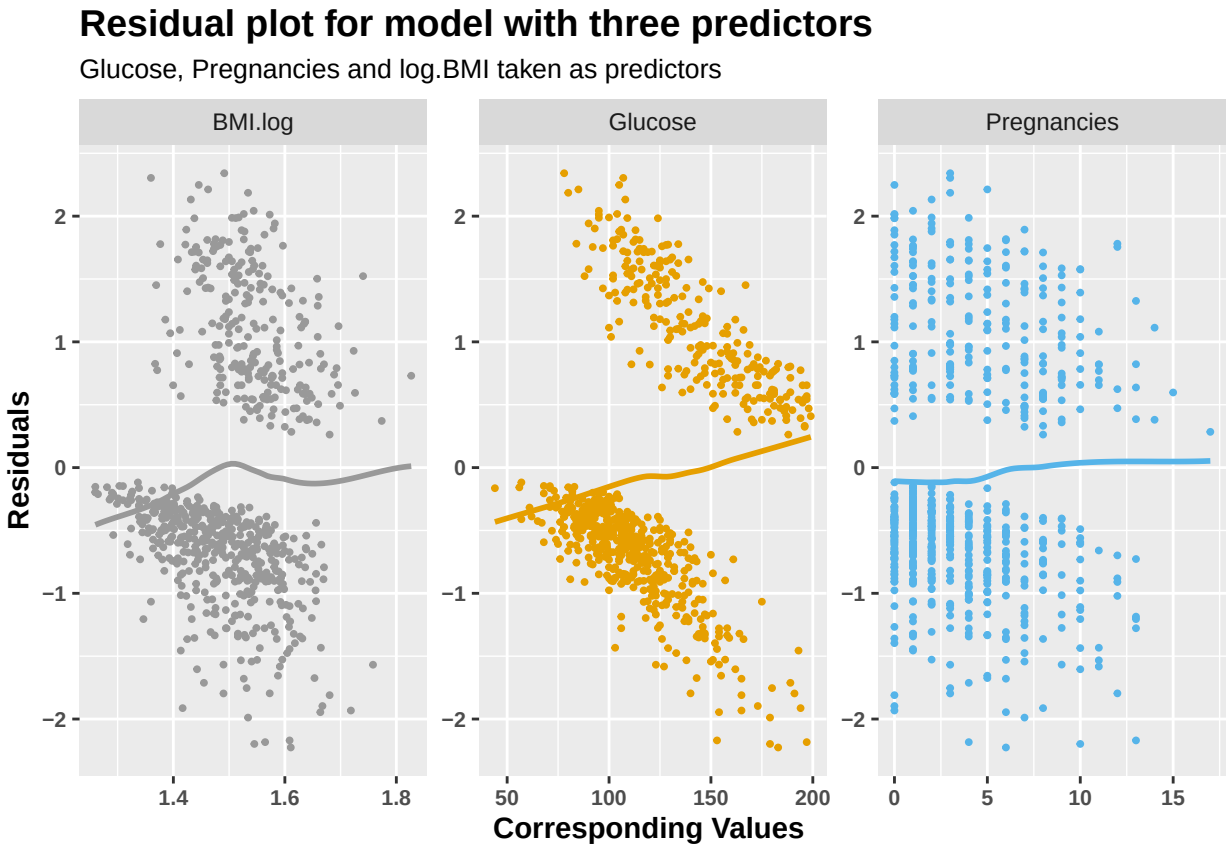
References

[1] : https://www.statista.com/topics/1723/diabetes/#dossierSummary___chapter1
[2] : https://www.who.int/news-room/fact-sheets/detail/diabetes
[3] : https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
[4] : Sturtevant WC. Handbook of North American Indians: History of Indian-White Relations. United States Govt Printing Office; 1983.
[5] : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4418458/
[6] : https://care.diabetesjournals.org/content/29/8/1866
[7] : https://www.ncbi.nlm.nih.gov/pubmed/7988310
[8] : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3968571/
[9] : https://www.health.harvard.edu/heart-health/reading-the-new-blood-pressure-guidelines
[10] : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3275515/ [11] : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1120973/

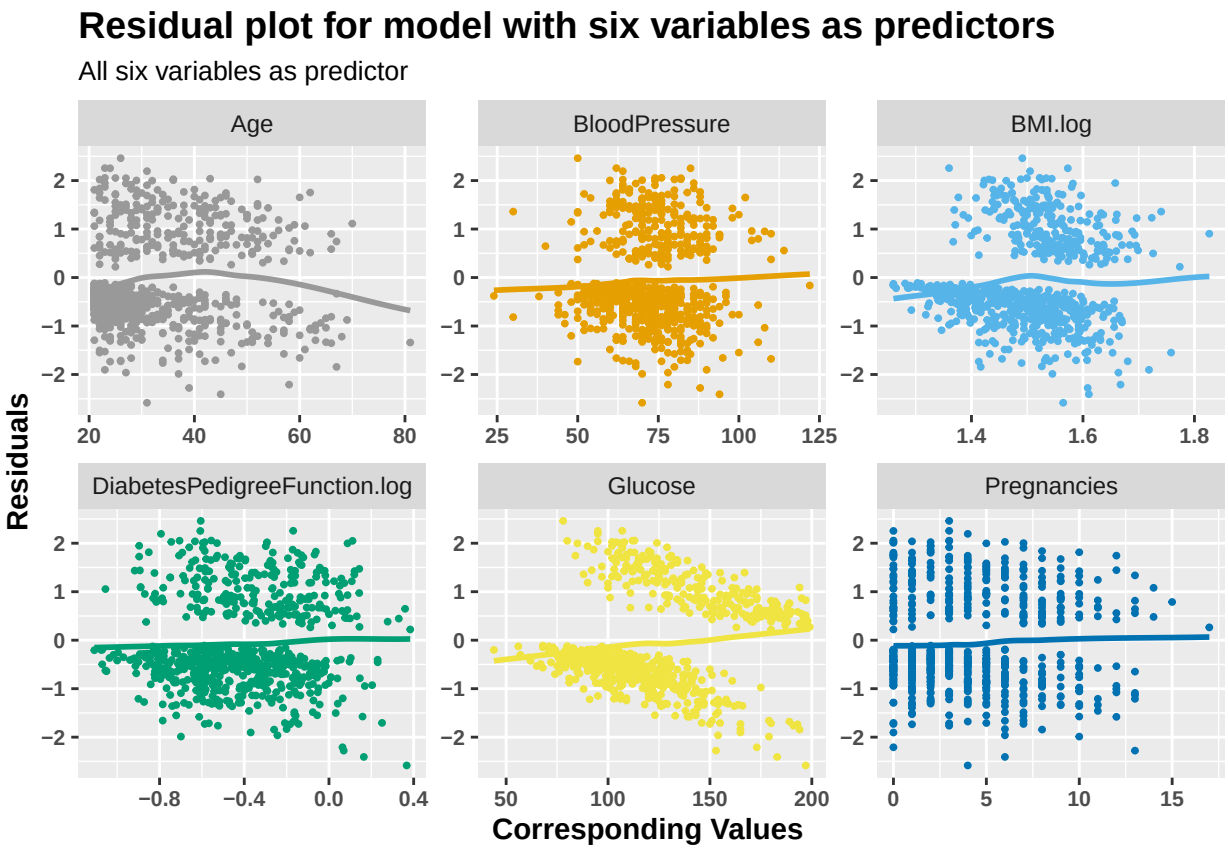
Appendix

Residual plots with a larger sample size.

Model with Glucose, log.BMI and Pregnancies are explanatory variables.



Model with all variables as predictor.



Residual plots with a smaller sample size.

Here, we show the residual plot with a smaller sample size of 392.

With Pregnancies, Glucose, BMI, and Insulin as predictors.