Analysing Diabetes of Pima Indian Women Using Logistic Regression Model

Shiyu Sun (UCID-30220704) Hyojin Kim (UCID-30222693)
Rukhsana Rashid Binti (UCID-30179306)

2023-12-10

Contents

1 Introduction

1.1 Background

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Diabetes is one of the most common human diseases and has become a significant public health concern worldwide. Over time, diabetes can damage blood vessels in the heart, eyes, kidneys and nerves. People with diabetes have a higher risk of health problems including heart attack, stroke and kidney failure. Diabetes can cause permanent vision loss by damaging blood vessels in the eyes. Many people with diabetes develop problems with their feet from nerve damage and poor blood flow. This can cause foot ulcers and may lead to amputation[1].

Type 2 diabetes is a multifactorial condition influenced by a combination of genetic, lifestyle, and environmental factors. Genetic predisposition plays a significant role, with individuals having a family history of diabetes being at higher risk. Lifestyle factors, such as sedentary behavior, poor dietary choices characterized by excessive consumption of refined sugars and saturated fats, and obesity, contribute significantly to the development of insulin resistance and impaired glucose metabolism. Aging is also a risk factor, as the body's ability to regulate blood sugar tends to diminish with age[2].

However, risk factors for one ethnic group may not be generalized to others. Being of Indigenous descent can increase the risk of living with type 2 diabetes [3]. Asian Indians have the highest diabetes prevalence rate (14.2%), whereas Asian Americans from Korea and Japan have the lowest diabetes prevalence rates 4.0% and 4.9%, respectively[4].

The prevalence of diabetes is reported to be higher among the Pima Indian community. The Pima are one of the tribes of American Indian and Alaska Natives and claim a population of approximately 20,000 members in 2010. They have the highest rate of type 2 diabetes in the world with 34.2% for Pima men and 40.8% for Pima women compared to 9.3% in the United States [5].

Therefore, this study used a dataset related to diabetes of Pima Indian Women to build a generalized linear model and investigate the significance of health-related predictors of diabetes in Pima Indians Women.

1.2 Motivation

In this comprehensive study, our focus centers on understanding the prevalent issue of diabetes among Pima Indian women, employing a meticulous examination of key health risk factors. The Pima community, exhibits one of the highest rates of type 2 diabetes globally, serves as a crucial backdrop for our investigation. Through statistical modeling, we aim to assess the individual significance of identified risk factors, such as BMI, glucose concentration, familial predisposition (Pedigree), age, and insulin levels, in influencing the likelihood of diabetes occurrence within this population.

Our analytical approach revolves around Generalized Linear Model, providing a nuanced understanding of the relationships between these health variables and diabetes prevalence. The study unfolds through hypothesis testing, confidence interval estimation, and model fitness assessments, enabling us to quantify the impact of each factor independently. Through revealing the distinct roles played by these predictors, our study aims to offer practical understanding of the factors influencing diabetes risk in the Pima Indian female population. By using this lens, our research hopes to make a significant contribution to the larger conversation on diabetes prevalence and health risk factors, which will ultimately lead to better informed healthcare interventions and community well-being.

2 Data Collection

The sample used in the analysis is derived from the Pima Indian Diabetes Dataset, originally provided by the National Institute of Diabetes and Digestive and Kidney Diseases [6] . The original dataset consists of 768 women residing near Phoenix, Arizona, USA. However, the dataset we use for analysis is from Harvard Dataverse [7], where missing values of specific columns have been removed, resulting a reduction of rows to 394. So our sample size is 394.

Our dataset includes 8 medical predictor variables such as time of pregnancies, glucose tolerance test, blood pressure, skin thickness, pedigree diabetes function, BMI, insulin levels, and age. Along with a binary response variable called "Diabetes".

2.1 Data Preprocessing

Default values for diabetes are -1 and 1. We changed them to 0 and 1 to be suitable as the response variable of the logistic regression model. Since pregnancies is discrete variable and in order to avoid plenty of dummy variables, we classified them to 5 categories, 0, 1, 2, 3, 4, and treated them as factors. This data preprocessing allows a more convenient model building and data analysis.

Below is a data dictionary describing the predictor variables used in the analysis:

Response variable: Diabetes (Binary, 0 = normal, 1 = diabetes)

Predictor Variable	**Format**	**Description**
Pregnancies group (pregnancies)	Discrete	The number of pregnancies classified into 5 groups (0, 1, 2, 3, 4) to avoid dummy variables.
Oral Glucose Tolerance Test (glucose)	Continuous	Two-hour plasma glucose concentration after a 75g anhydrous glucose test in mg/dl.
Blood Pressure (BP)	Continuous	Diastolic Blood Pressure in mmHg.
Skin Thickness (thickness)	Continuous	Triceps skin fold thickness in mm.
Insulin (insulin)	Continuous	2-hour serum insulin in mu U/ml.
BMI (BMI)	Continuous	Body Mass Index calculated as weight in kg divided by the square of height in meters.
Pedigree Diabetes Function (pedigree)	Continuous	A function representing the likelihood of getting the disease based on ancestral history.
Age (age)	Continuous	Age in years.

Table 1: Description of Predictor Variables

2.2 Descriptive Data Analysis

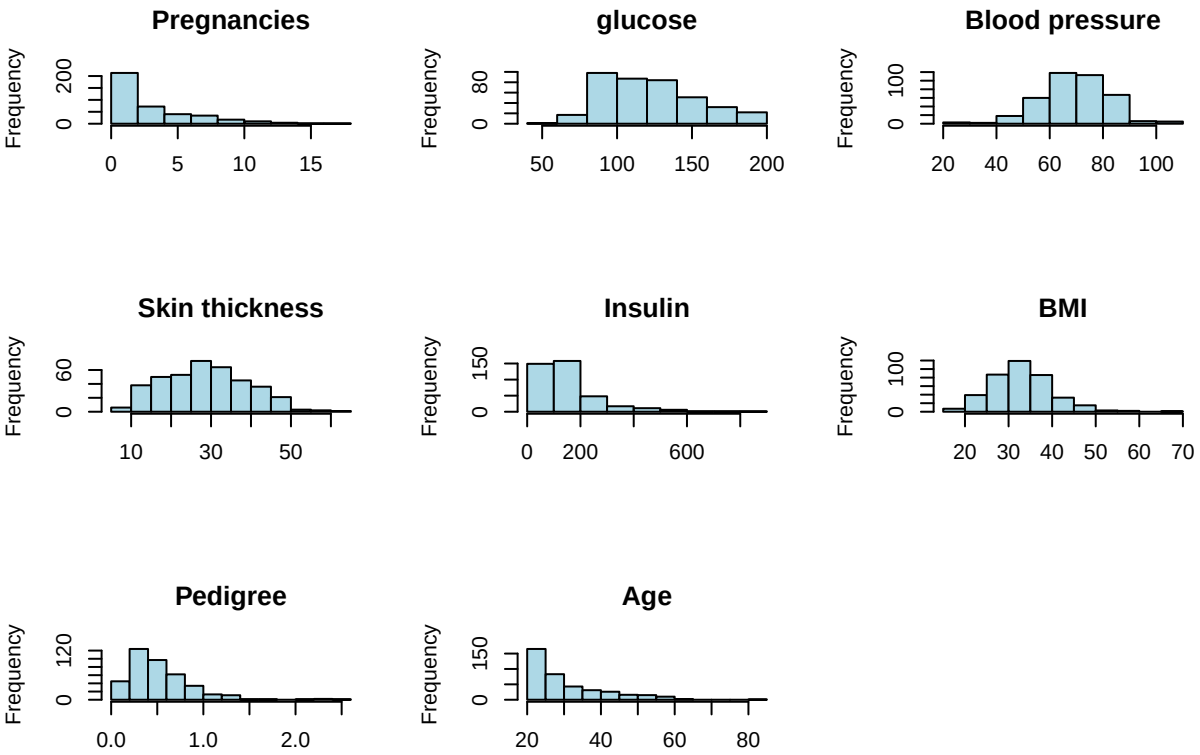
Variable	**Minimum**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
pregnancies	0	1	2	3.30	5	17
glucose	0.00	99.00	119.00	122.38	143.00	198.00
BP	24.00	62.00	70.00	70.67	78.00	110.00
thickness	7.00	21.00	29.00	29.13	36.75	63.00
insulin	0.00	76.00	125.00	155.32	190.00	846.00
BMI	18.20	28.40	33.20	33.07	37.07	67.10
pedigree	0.085	0.27	0.45	0.52	0.68	2.42
age	21	23	23	30.88	36	81

Table 2: Summary statistics of Variables

The summary statistics (minimum, 1st quartile, median, mean, 3rd quartile and maximum) of the dataset Pima Indian Diabetes is given above.

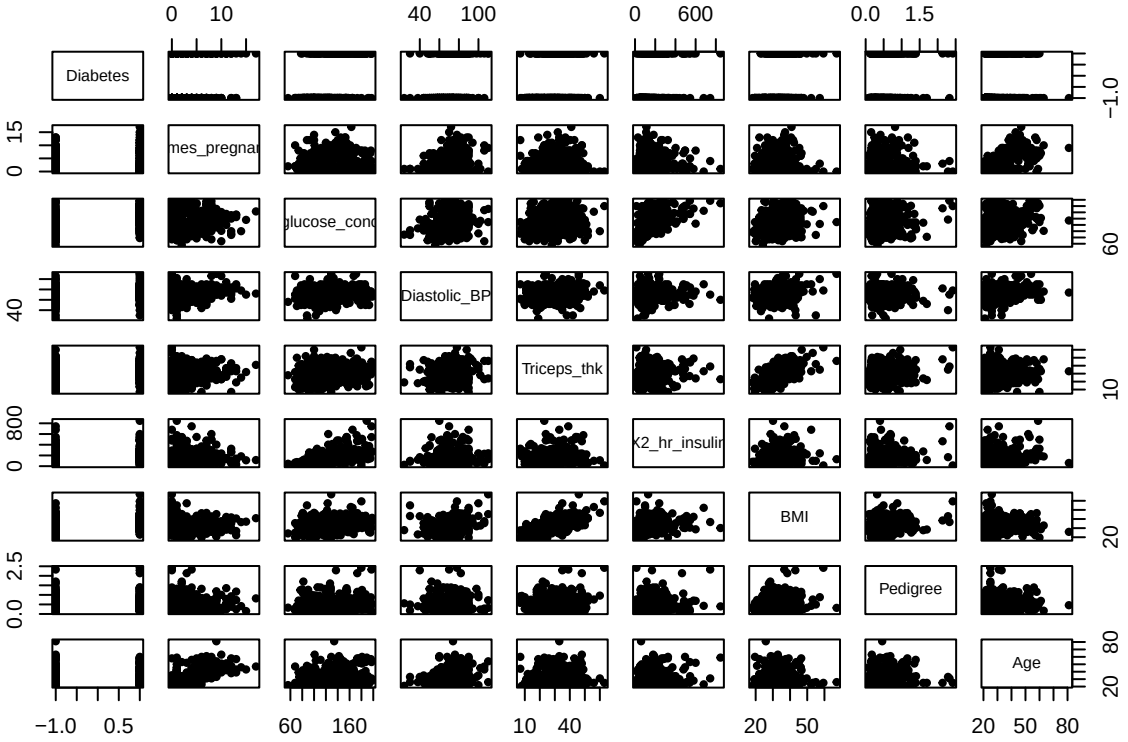
We noticed there are some "0"'s for glucose and insulin, they are outliers and we removed these rows from the dataset.

1.Following is the histograms of our response variable and 8 predictor variables.



From the histograms, we can see the distribution of pregnancies, glucose, insulin, pedigree and age can be seen as right skewed. While BP, thickness, BMI and pedigree approximately apply to normal distribution.

2. Analysis on pairs of variables



From the scatter plot, we noticed that there is a positive correlation between BMI and skin thickness, pregnancies and age. As there is a positive trend for the data points in these two scatter plot.

2.3 Statistical Model Choosing

Following is the values of our response variable "Diabetes"

$$\text{Diabetes} = \begin{cases} 1 & \text{if the patient has diabetes} \\ 0 & \text{if the patient does not have diabetes} \end{cases}$$

As it is binary, we are going to use logistic regression model".

Logistic Regression Model:
Logistic Regression is a statistical method used for modeling the probability of a binary outcome meaning it takes on two possible values, often coded as 0 and 1. This could represent outcomes such as 'success' or 'failure,' 'yes' or 'no,' or 'positive' or 'negative.'

For Logistic regression model we need to use a link function called logit function.
The logit function in logistic regression is defined as:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where β_i is the coefficient, and X_i is the predictor variable. We could also add interaction terms based on hypothesis testing and select the best model among our candidate models.

3. Hypothesis Testing

In statistical models, hypothesis testing for coefficient values is essential to verify the accuracy of predictions and evaluate how well the model fits the data. The weights of the independent variables are represented by beta values, and the statistical significance of these values shows how significant the relationships they depict are. This procedure supports decision-making for model improvement, including the inclusion or removal of variables, and validates the model's performance in interpreting the findings. In general, hypothesis testing makes statistical models more reliable and comprehensible, resulting in a stronger comprehension of the correlations between the variables.

Furthermore, through hypothesis testing, we can discern which variables exhibit a significant relationship with diabetes. By assessing the statistical significance of the beta values associated with each variable, we gain insights into their individual contributions to the prediction of diabetes. This information aids in the identification of key factors influencing diabetes, guiding further model refinement and enhancing our understanding of the complex relationships within the dataset. Overall, the iterative process of hypothesis testing not only ensures the robustness of the model but also allows us to unravel meaningful associations that contribute to a comprehensive understanding of diabetes determinants.

3.1 Hypothesis Test of Original Model

In this project, the original model, which initially involved one-way single variables, underwent fitting. A Wald test was subsequently employed to determine the significance of variables and identified any non-significant ones.

Below shows the hypothesis and the result of test.
H_0: The coefficient of each explanatory variable is zero.
H_1: The coefficient of each explanatory variable is not zero.

Variable	z-value	p-value
(Intercept)	-7.46	8.30e-14
glucose	6.48	9.04e-11
BP	-0.17	0.86
thickness	0.77	0.44
insulin	-0.71	0.48
BMI	2.42	0.015
pedigree	2.64	0.0082
age	2.42	0.015
pregnant1	-0.44	0.66
pregnant2	-0.11	0.92
pregnant3	0.44	0.66
pregnant4	0.36	0.72

Table 3: z-statistic and p-value of original model

From the results of the Wald test, only five coefficients, including the intercept, were statistically significant at the 5% significance level. (Intercept), glucose, BMI, pedigree, and age appeared to be significant predictors, whereas BP, thickness, insulin, and the pregnancies variables lack significance in the model. This outcome prompted consideration of potential interaction terms influencing diabetes.

The absence of statistical significance for certain variables suggested that their individual effects may not be reliable indicators. However, it underscored the possibility that interactions among variables collectively impact diabetes prediction. Further exploration of interaction terms could enhance the model's explanatory power, offering a more profound understanding of the variables influencing diabetes in this research.

3.2 Hypothesis Test of Model with Two-Way Interaction Terms

Following the test results of the initial model, we advanced to fitting an extended model that integrates two-way interaction terms.
Diabetes ~ (glucose+BP+thickness+insulin+BMI+pedigree+age+pregnancies)ˆ2
This augmented model preserved the original eight single variables while introducing all conceivable two-way interaction terms. It will be referred to as the "two-way model."

We performed the Wald-test using the same hypothesis using the two-way model. Similar to the hypothesis testing results of the original model, the two-way model also had that only a few variables are statistically significant. The significant variables are shown below table.

Variable	z-value	p-value
pregnant1	-1.99	0.047
glucose:BMI	-2.64	0.0082
insulin:pedigree	-2.32	0.020
age:pregnant1	2.64	0.0082

Table 4: Significant variables of two-way interaction model from Wald Test

3.3 Cutting Down Method

The two-way model contains numerous variables that are not statistically significant, and the overall size of the model is excessively large. In an effort to streamline the model and retain only statistically significant variables, we conducted hypothesis testing once again.

To remove the non-significant two-way interaction variables, we tested Likelihood Ratio Test(LRT) using the drop1() function in R. LRT evaluates the change in model fit by removing a single variable to the model.

The hypothesis are as follows:
H_0: The removed variable does not contribute significantly to the model, resulting in no significant change in model fit.
H_1: The removed variable contributes significantly to the model, leading to a significant change in model fit.

By LRT test, we could get five statistically significant two-way interaction terms: 1) glucose:BMI, 2) thickness:age, 3) insulin:pedigree, 4) insulin:age, and 5) Age:pregnancies.

Variable	z-value	p-value
glucose:BMI	7.34	0.0067
thickness:age	3.93	0.047
insulin:pedigree	6.80	0.0091
insulin:age	3.97	0.046
age:pregnancies	9.55	0.049

Table 5: Significant variables of two-way interaction model from LRT test

We proceeded to fit a model by retaining only these five significant interaction terms and excluding the others. This model included the original eight variables from the original model, along with the five two-way interaction terms identified as significant in the test results.
Diabetes ~ glucose+BP+thickness+insulin+BMI+pedigree+age+pregnancies+glucose:BMI+thickness:age +insulin:pedigree+insulin:age+age:pregnancies

From this model we applied the Wald test to each of the model's coefficients and removed the most non-significant variable.
H_0: The coefficient of each explanatory variable is zero.
H_1: The coefficient of each explanatory variable is not zero.

Similar to the previous hypothesis test results, this model includes many variables that are not statistically significant. Among them, the least significant variable is BP and its p-value is 0.7031.

From the above result, BP variable was removed from the model and fit the data again.
Diabetes ~ glucose+thickness+insulin+BMI+pedigree+age+pregnancies+glucose:BMI+thickness:age +insulin:pedigree+insulin:age+age:pregnancies

The hypothesis test is perforemd again and the variable found to be the least significant is age, with a p-value of 0.364. In the subsequent analysis, the decision was made to conduct a test by removing the age variable.

Similar to the hypothesis testing conducted earlier, the iterative process of systematically removing variables is continued. This sequential approach aims to eliminate non-significant variables, one step at a time, until only statistically significant variables remain in the model. This method allows for a refined model that emphasizes the inclusion of predictors with a meaningful impact on the response variable, contributing to a more streamlined and interpretable statistical model.

Continuing this iterative process of variable removal, the final model that emerged is represented by the following formula:
Diabetes ~ glucose + insulin + BMI + pedigree + insulin:pedigree + insulin:age
This refined model includes only the statistically significant variables, such as glucose concentration, insulin, BMI, pedigree, and their interactions, specifically insulin with pedigree and age.

Variable	z-value	p-value
(Intercept)	-8.78	< 2.22e-16
glucose	7.19	6.07e-13
insulin	-2.39	0.016
BMI	3.80	0.00014
pedigree	3.95	7.76e-05
insulin:pedigree	-2.79	0.0051
insulin:age	4.14	3.39e-05

Table 6: z statistic and p-value of the final model

3.4 Adding Up Method

We also did in another way, by adding interaction terms to get the most suitable model. Model 1 represents estimates from the null model, while Model 2 is our original model. Since the residual deviance is decreased significantly with the inclusion of predictor variables, we can say that the model with the inclusion of predictor variables is better than the null model. Based on our original model, the variables pregnancies, thickness, BP, and insulin are not significant predictors of diabetes at 5% significance level.

Model 3 is the reduced model with only 4 predictor variables, BMI, glucose, pedigree, and age. Non-significant variables from Model 2 are omitted, as we want to ensure every variable is significant, providing more meaningful result.

Next, we added interaction terms based on the reduced model to improve goodness of fit. We started with adding 7 interactions of first variable (BMI) with other 7 variables to Model 3, and then analyse the significance of all variables (singles and interactions). 8 additional models with different interactions were built and analysed in this way. Among these models, we observed from the one including interactions related insulin, that insulin interacting with pedigree and age is significant, while single variable age is not. Therefore, age is removed and insulin (this single variable turns to be significant in the model) is added. This is the steps of getting Model 4, the final model, which consists of BMI, glucose, pedigree, insulin, insulin $\times$ age, and insulin $\times$ pedigree.

Following is the table of estimates of 4 models.

Figures in parentheses indicate p-values. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.1$. BP and thickness represent blood pressure and skin thickness.

	Model 1	Model 2	Model 3	Model 4
intercept	$-0.70_{(***)}$	$-1.00_{(***)}$	$-10.13_{(***)}$	$-1.03_{(***)}$
BMI		$-0.07_{(*)}$	$0.08_{(***)}$	$0.08_{(***)}$
pedigree		$1.13_{(**)}$	$1.09_{(**)}$	$2.61_{(***)}$
glucose		$-0.04_{(***)}$	$0.04_{(***)}$	$0.04_{(***)}$
age		$0.05_{(*)}$	$0.05_{(***)}$	
insulin		0.0009		$0.007_{(*)}$
BP		-0.0021		
thickness		0.013		
pregnancies1		0.21		
pregnancies2		-0.056		
pregnancies3		0.21		
pregnancies4		0.20		
pedigree:insulin				$-0.006_{(**)}$
age:insulin				$0.0009_{(***)}$

Table 7: Estimated coefficients of 4 candidate models

Above all, although we used two different methods of cutting down and adding up interactions, the final model is the same.

The final model is
$\text{logit}(diabetes) = -10.2666 + 0.0436x_{glucose} - 0.00697x_{insulin} + 0.08242x_{BMI} + 2.6093x_{pedigree} - 0.0062x_{insulin:pedigree} + 0.0003x_{insulin:Age}$

Each coefficient including the intercept can be interpreted that:
(Intercept): The estimated intercept is approximately -10.27. This represents the log-odds of the response variable when all other predictors are zero.
glucose: For every one-unit increase in glucose concentration, the log-odds of the response variable increase by approximately 0.043.
insulin: For every one-unit increase in X2_hr_insulin, the log-odds of the response variable decrease by approximately 0.007.
BMI: For every one-unit increase in BMI, the log-odds of the response variable increase by approximately 0.082.
pedigree: For every one-unit increase in Pedigree, the log-odds of the response variable increase by approximately 2.61.
insulin:pedigree: The interaction term between insulin and pedigree. For every one-unit increase in this interaction term, the log-odds of the response variable decrease by approximately 0.006.
insulin:age: The interaction term between insulin and age. For every one-unit increase in this interaction term, the log-odds of the response variable increase by approximately 0.00029.

4. Confidence Interval

We have calculated confidence intervals for the regression coefficients in the final model chosen using hypothesis testing and the original model. Each independent variable's confidence intervals sheds light on how consistently the effects connected to each variable are reliable. These intervals improve our comprehension of the relative effects of variables and aid in evaluating the model's stability.

In statistical terms, a confidence interval represents a range of values within which we are reasonably confident that the true parameter lies. A 95% confidence interval implies that if we are approximately 95% confident that the true coefficient value is in between CI.

These confidence intervals, which provide an indicator of the accuracy and consistency of the predicted coefficients, are essential to understanding the model. They contribute to a deeper comprehension of the

connections inside the model.

95% Confidence interval: (Coefficients estimation $\pm$ 1.96 $\times$ Coefficient standard error)

4.1 Confidence Interval of Original Model

The 95% of confidence intervals of original model coefficients are calculated.

Variable	Estimates	SE	Lower	Upper
(Intercept)	-10.01	1.34	-12.63	-7.38
glucose	0.03	0.0059	0.026	0.049
BP	-0.0021	0.012	-0.026	0.021
thickness	0.013	0.017	-0.020	0.047
insulin	-0.00093	0.0013	-0.0035	0.0016
BMI	0.067	0.027	0.012	0.12
pedigree	1.13	0.43	0.29	1.97
age	0.046	0.019	0.0087	0.083
pregnant1	-0.21	0.49	-1.16	0.74
pregnant2	-0.056	0.52	-1.092	0.98
pregnant3	0.21	0.48	-0.46	1.47
pregnant4	0.20	0.57	-0.54	1.69

Table 8: 95% CI of original model coefficients

Similar to the hypothesis result, some variables include zero in the confidence interval. This means that the variables which include zero in the confidence interval, does not statistically significant.

4.2 Confidence Interval of the Final Model

Variable	Estimates	SE	Lower	Upper
(Intercept)	-10.27	1.17	-12.56	-7.98
glucose	0.043	0.0061	0.032	0.055
insulin	-0.0069	0.0029	-0.013	-0.0013
BMI	0.08	0.022	0.040	0.12
pedigree	2.61	0.66	1.32	3.90
insulin:pedigree	-0.0061	0.0022	-0.01	-0.0018
insulin:age	0.00029	0.000071	0.00015	0.00043

Table 9: 95% CI of final model coefficients

The 95% of confidence intervals of the final model coefficients are calculated.

The final model's confidence interval contains no variables that include 0, in contrast to the original model. We are able to confirm that every variable is important.

5 Model Checking

In the realm of statistical modeling, assessing the goodness of fit is a critical step to ensure that a chosen model aligns well with the observed data. One widely employed metric for this purpose is the coefficient of determination, commonly known as R^2. In this case, McFadden's pseudo R^2 is calculated to check the model. This metric quantifies the proportion of variability in the dependent variable that is explained by the model, providing valuable insights into the model's effectiveness in capturing the underlying patterns within the data.

5.1 McFadden's Pseudo R-squared Test

The equation of McFadden's Pseudo R^2 is as following

$$R^2 = \frac{l(b_{min}) - l(b)}{l(b_{min}) - l(b_{max})} = \frac{Dev(b_{min}) - Dev(b)}{Dev(b_{min})}$$

Below is R^2 value of our four candidate models. Model 1 represents the null model, model 2 is the original model, model 3 is the reduced model with 4 single variables, while model 4 is our final model.

	Model 1	Model 2	Model 3	Model 4
McFadden's Pseudo R^2	1.13e-16	0.310	0.306	0.338

Table 10: R^2 values

Based on the computed pseudo R^2 values, the R^2 of original model (model 2) is 0.31, indicating that it explains approximately 31.06% of the variability in the dependent variable. This value suggests a moderate level of explanatory power. In contrast, the final model exhibits an improved R^2 of 0.34, representing a higher percentage of explained variability.

If the R^2 value ranges from 0.2 to 0.4, then it indicates a good model fit, thus model 2, 3, 4 are all well fitted models compared to the null model.

5.2 Deviance Test

The deviance-based goodness-of-fit test is an essential technique for evaluating the quality of a model in statistical modeling. Deviance, a measure of how well the model fits the observed data, is the primary focus of this test. Deviance statistics analysis supports in detecting potential model flaws and advances efforts to make the model more accurate and reliable.

The following is the hypothesis that was used in the goodness of fit test.
H_0: The model fits the data well.
H_1: The model does not fit the data well.

Deviance statistic is denoted as G^2, it is also called the likelihood-ratio test (LRT) statistic in some text. The formula is as follows.

$$G^2 = 2 \times (l(b_{max}) - l(b))$$

where $l(b_{max})$ is the log-likelihood of saturated model, while $l(b)$ is the log-likelihood of proposed model.

The deviance residual and corresponding p-value of our four models are given below.

	Model 1	Model 2	Model 3	Model 4
Deviance statistic G^2	501.11	345.46	347.81	331.73
p-value	0.00017	0.91	0.93	0.98

Table 11: Deviance statistic and corresponding p-value of candidate models

For Model 1, the p-value is 0.00017, less than our significance level 0.05. It gives us sufficient evidence to reject null hypothesis, indicating the null model does not fit with our data.

Model 2, Model 3, Model 4 all have lower deviance statistics (G^2) compared to Model 1. Their corresponding p-values are much larger than 0.05. Therefore, we have insufficient evidence to reject null hypothesis for

these three models. The goodness-of-fit tests for Model 2, 3, 4 are not significant, they are good fit for our data.

6 Model Comparison and Selection

For this section, we compared our four candidate models to obtain best possible model.

Since they are not nested, likelihood-ratio test and deviance test could not be used to perform the comparison, which are restricted to be only valid for nested models. We used Akaike's Information Criteria (AIC), and Bayesian Information Criteria (BIC) to select the best possible model.

6.1 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) serves as an estimator of prediction error and facilitates model selection by assessing the relative quality of different statistical models applied to a given dataset. It acknowledges that statistical models never perfectly represent the underlying data-generating process, which results in some information loss. AIC quantifies this information loss, offering a measure to compare models. When fitting models, it is possible to increase the maximum likelihood by adding parameters, but doing so may result in overfitting. This criterion effectively balances the trade-off between a model's goodness of fit and its simplicity, dealing with risks of both overfitting and underfitting.

Let p be the number of parameters, while $\hat{L}$ is the maximized value of the likelihood function. For GLM, $-2\log(\hat{L})$ is the deviance of the model.

The AIC value of the model is as following

$$\text{AIC} = -2\log(\hat{L}) + 2p$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.

	Model 1	Model 2	Model 3	Model 4
AIC	503.11	369.46	357.81	345.73

Table 12: AIC values of candidate models

Consider the differences in AIC between models. If the difference is substantial (greater than 2), it suggests that the model with the lower AIC is significantly better.

As can be seen from the Table, AIC decreases from Model 1 to Model 4, indicating an improvement in model fit. Model 4 has the lowest AIC, which is favored by AIC as it provides the best compromise between goodness of fit and simplicity.

There will almost always be information lost due to using a candidate model to represent the "true model",and we hope to select the model that minimizes the information loss among the candidate models.

Denote the AIC value of our 4 models are AIC_1,AIC_2,AIC_3,AIC_4. Let AIC_{min} be the minimum of those values, which is $\text{AIC}_4 = 345.7268$. $\exp((\text{AIC}_{min} - \text{AIC}_i)/2)$ is known as the relative likelihood of model i, it can be interpreted as being proportional to the probability that the i^{th} model minimizes the information loss.

	Model 1	Model 2	Model 3
$\exp((\text{AIC}_{min} - \text{AIC}_i)/2)$	$6.68e^{-35}$	$7.02e^{-6}$	0.0024

Table 13: Relative likelihood of candidate models

From this Table, Model 3 is 0.00237 times as probable as Model 4 to minimize the information loss. For Model 1 and Model 2, they are much less likely to minimize the information loss compared to Model 4.

Above all the comparison analysis using AIC, Model 4 is chosen to be the best model. While AIC is useful for model selection, we also need BIC to assess if Model 4 is also preferred by this criteria.

6.2 Bayesian information criteria (BIC)

The Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred. It is based on the likelihood function and it is closely related to the Akaike information criterion (AIC).

Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model. Note that BIC incorporates a higher penalty for a higher number of observations, and so it rewards more parsimonious models. The penalty for model complexity increases with the sample size, making BIC more stringent for larger datasets.

Similarly, p is the number of parameters, $\hat{L}$ is the maximized value of the likelihood function estimate. For GLM, $-2\log(\hat{L})$ is the deviance of the model. n is the number of observations.

The BIC value of the model is as following

$$\text{BIC} = -2\log(\hat{L}) + p \times \log(n)$$

We calculated the BIC of each candidate model and got the following table.

	Model 1	Model 2	Model 3	Model 4
BIC	507.09	417.17	377.69	373.56

Table 14: BIC values of candidate models

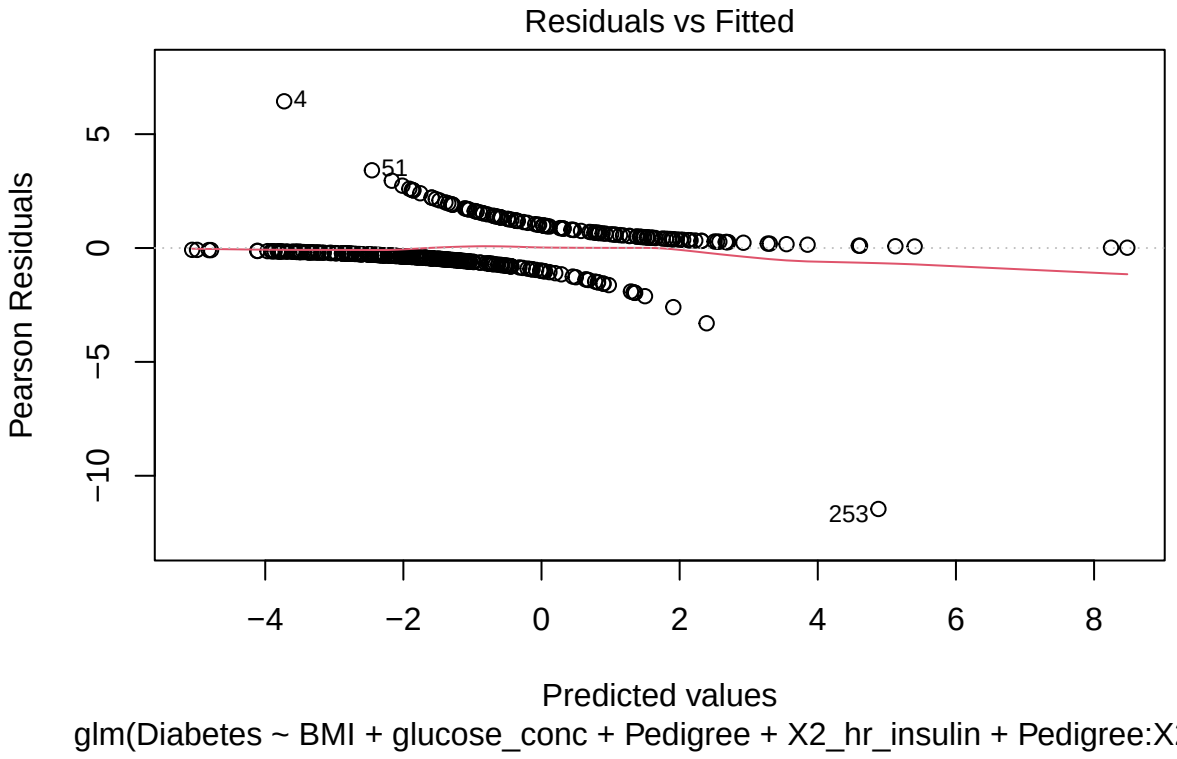
When picking from several models, ones with lower BIC values are generally preferred. From the table, we can see that Model 4 still has the lowest BIC value, indicating it is also favored by BIC among four candidate models.

In conclusion, we select Model 4 as the best model for our dataset. It is preferred by both the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC). This model reached a balance between model fit and complexity, and demonstrates a strong capacity to capture underlying patterns in the data while avoiding unnecessary intricacies that might lead to overfitting. The agreement between AIC and BIC suggests that the selected model is both parsimonious and effective in explaining the observations.

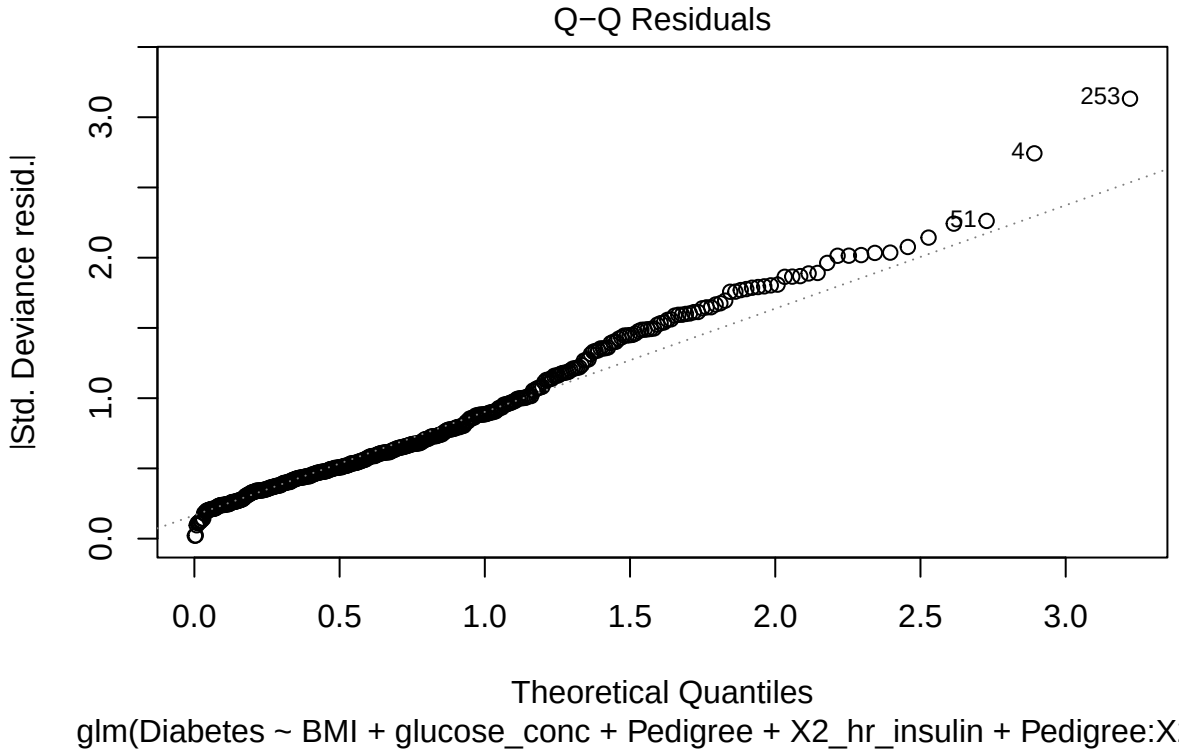
7 Plots Interpretation

In this section, we are going to interpret the Residuals vs Fitted plot, Q-Q plot of Model 4, our best fitted model.

1. From Residuals vs Fitted plot for Model 4, we can see there are two curvilinear trends, since the fit of a logistic regression is curvilinear by nature. The red line is the estimated regression line, and data points 4, 51 and 253 deviate from it. Since these three points are far from the main cluster, they are identified as outliers.



2. From the Normal Q-Q plot, it is shown that our residuals are normally distributed except the upper tail of observations, where the outliers are also included. For the data in upper tail, the deviance residuals are higher than theoretical one, they are more spread out than supposed.



8 Conclusion

We analyzed the Pima Indian Diabetes dataset to identify variables associated with diabetes through hypothesis testing. We built several logistic regression models using statistically significant variables. We then calculated each model by calculating R^2 and performed deviance tests to measure how well the model fit the data.

Comparing models using AIC and BIC informs decisions about which model best fits the data. Through this comparison, We expanded our understanding of the model's fitness and observed improved performance as the model was expanded.

Our project provided a deep understanding of hypothesis testing, logistic regression modeling, and application of model evaluation metrics. Insights gained from comparing and evaluating different models will likely prove invaluable in future analyzes of similar data sets. This project not only deepened our understanding of concepts in statistics and modeling, but also gave us practical experience in data analysis.

9 References

References

[1] Krasteva A., Panov V., Kisselova A., Krastev Z. Oral cavity and systemic diseases—Diabetes mellitus. Biotechnol. Biotechnol. Equip. 2011;25:2183–2186. doi: 10.5504/BBEQ.2011.0022.

[2] Tremblay, J., Hamet, P. (2019). Environmental and genetic contributions to diabetes. Metabolism, 100, 153952. https://doi.org/10.1016/j.metabol.2019.153952

[3] "Statistics about Diabetes." Statistics About Diabetes | ADA, diabetes.org/about-diabetes/statistics/about-diabetes. Accessed 03 Dec. 2023.

[4] Spanakis, Elias K., and Sherita Hill Golden. 'Race/Ethnic Difference in Diabetes and Diabetic Complications'. Current Diabetes Reports, vol. 13, no. 6, Dec. 2013, p. 10.1007/s11892-013-0421–29. PubMed Central, https://doi.org/10.1007/s11892-013-0421-9.

[5] Booth, Clayton. Policy and Social Factors Influencing Diabetes among Pima Indians in Arizona, USA. Public Policy and Administration Research, No 2017.

[6] Kahn,Michael. Diabetes. UCI Machine Learning Repository. https://doi.org/10.24432/C5T59G.

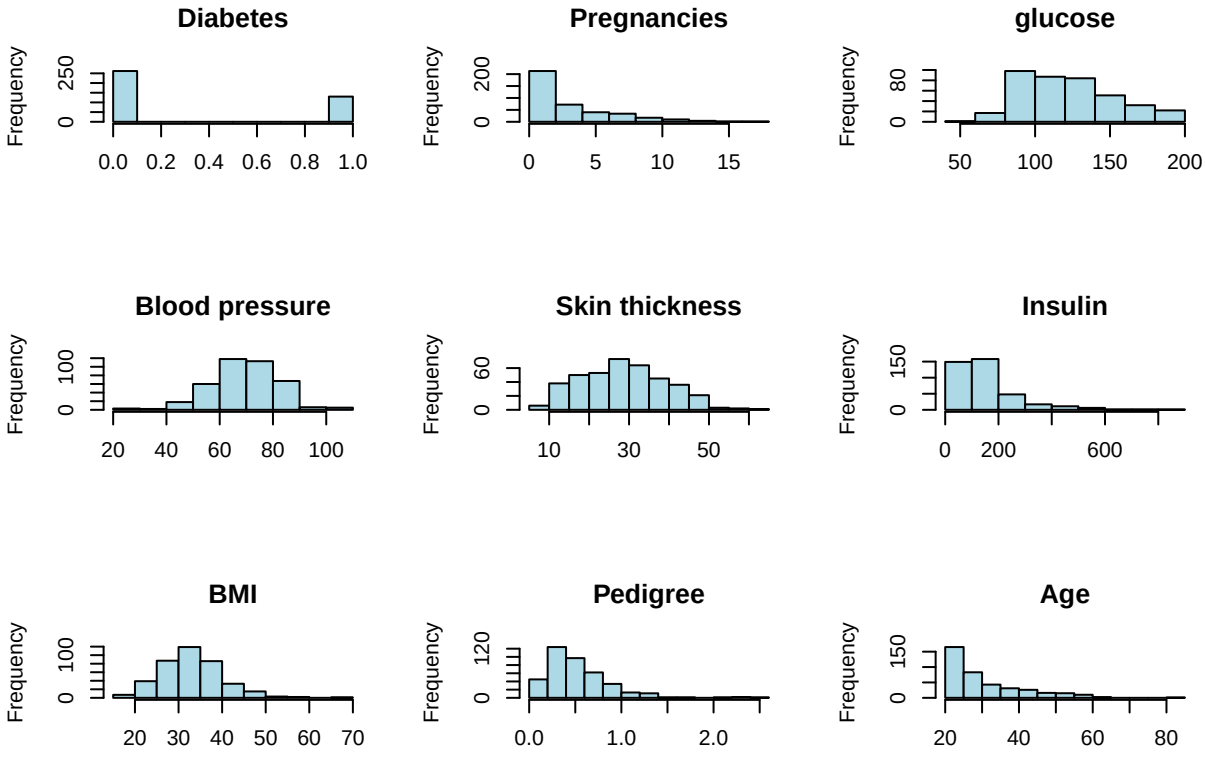
[7] Bartley, Christopher. Replication Data for: Pima Indians Diabetes. Harvard Dataverse. https://doi.org/10.7910/DVN/XFOZQR

10 Appendix

10.1 Descriptive Plots

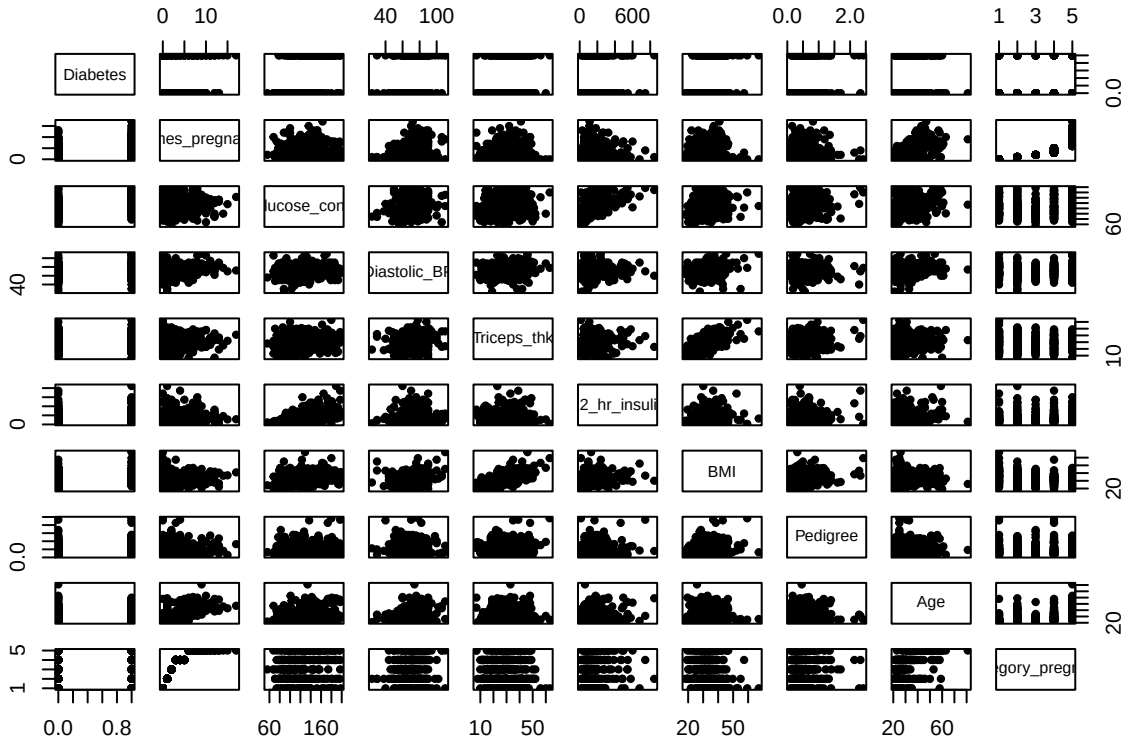
The histograms of response variable and predictor variables

```
db <- read.csv("original_data.csv",header=TRUE)
# Assuming your data frame is named 'db'
db <- subset(db, X2_hr_insulin != 0 & glucose_conc != 0)
# remove rows with insulin or glucose is equal to 0
db$Diabetes <- ifelse(db$Diabetes == -1, 0,1)  ## update the diabetes to 0,1
plot_data <- data.frame(db$times_pregnant,db$glucose_conc,db$Diastolic_BP,db$Triceps_thk,db$X2_hr_insuli
db$category_pregnant <- ifelse(db$times_pregnant ==0, 0,
                       ifelse(db$times_pregnant ==1, 1,
                       ifelse(db$times_pregnant ==2, 2,
                              ifelse((db$times_pregnant <=5)&(db$times_pregnant >=3),3,4))))  ## add col
db$category_pregnant <- as.factor(db$category_pregnant)  ## make pregnant category to factor values
par(mfrow = c(3, 3))
hist(db$Diabetes, col = "lightblue", main = "Diabetes",xlab = " ")
hist(db$times_pregnant, col = "lightblue", main = "Pregnancies",xlab = " ")
hist(db$glucose_conc, col = "lightblue", main = "glucose",xlab = " ")
hist(db$Diastolic_BP, col = "lightblue", main = "Blood pressure",xlab = " ")
hist(db$Triceps_thk, col = "lightblue", main = "Skin thickness",xlab = " ")
hist(db$X2_hr_insulin, col = "lightblue", main = "Insulin",xlab = " ")
hist(db$BMI, col = "lightblue", main = "BMI",xlab = " ")
hist(db$Pedigree, col = "lightblue", main = "Pedigree",xlab = " ")
hist(db$Age, col = "lightblue", main ="Age",xlab = " ")
```



Scatter plot for pairs of data points

```r
plot(db,pch=20)
```



10.2 Hypothesis Testing

To perform the 95% Wald test for the coefficients, first we fit the original model which has 8 single variables on the data. From the summary, we could check the result of Wald test.

```r
#Fit the original model
model2 <- glm(Diabetes ~ . - times_pregnant , data = db, family = binomial)
#Check p-value for Wald test
summary(model2)
```

```
##
## Call:
## glm(formula = Diabetes ~ . - times_pregnant, family = binomial,
##     data = db)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -9.9720905  1.3393539  -7.445 9.66e-14 ***
## glucose_conc        0.0375336  0.0058371   6.430 1.27e-10 ***
## Diastolic_BP       -0.0019042  0.0119705  -0.159  0.87361
## Triceps_thk         0.0132528  0.0172498   0.768  0.44232
## X2_hr_insulin      -0.0008506  0.0013204  -0.644  0.51945
## BMI                 0.0672096  0.0278289   2.415  0.01573 *
## Pedigree            1.1273580  0.4282093   2.633  0.00847 **
## Age                 0.0453897  0.0189270   2.398  0.01648 *
## category_pregnant1 -0.2110060  0.4880488  -0.432  0.66549
## category_pregnant2 -0.0566470  0.5283218  -0.107  0.91461
## category_pregnant3  0.2142258  0.4782555   0.448  0.65420
## category_pregnant4  0.1970059  0.5693556   0.346  0.72933
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.1  on 391  degrees of freedom
## Residual deviance: 345.0  on 380  degrees of freedom
## AIC: 369
##
## Number of Fisher Scoring iterations: 5
```

From the result of the original model, only few variables are significant. It made us to consider there might be some interaction terms influencing the diabetes. Therefore, we fit two-way interaction model and test the coefficient. From the summary function, we could check the result, and only some variables are significant, similar to the original model.

```
#Fit two-way interaction model
model5 <- glm(Diabetes ~ (glucose_conc+Diastolic_BP+Triceps_thk+X2_hr_insulin+BMI+Pedigree+Age+category_
#Check p-value for Wald test
summary(model5)
```

```
##
## Call:
## glm(formula = Diabetes ~ (glucose_conc + Diastolic_BP + Triceps_thk +
##     X2_hr_insulin + BMI + Pedigree + Age + category_pregnant)^2,
##     family = binomial, data = db)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.423e+01  1.216e+01  -1.170  0.24195
## glucose_conc                     1.260e-01  7.725e-02   1.631  0.10283
## Diastolic_BP                    -1.876e-01  1.635e-01  -1.147  0.25121
## Triceps_thk                      1.197e-01  2.214e-01   0.541  0.58871
## X2_hr_insulin                    5.435e-04  2.204e-02   0.025  0.98032
## BMI                              4.294e-01  2.984e-01   1.439  0.15015
## Pedigree                         8.276e+00  5.242e+00   1.579  0.11433
## Age                             -2.608e-01  3.041e-01  -0.857  0.39118
## category_pregnant1              -1.352e+01  6.679e+00  -2.025  0.04289 *
## category_pregnant2              -8.317e+00  6.617e+00  -1.257  0.20883
## category_pregnant3               2.711e-01  5.151e+00   0.053  0.95803
## category_pregnant4              -6.698e-01  6.085e+00  -0.110  0.91235
## glucose_conc:Diastolic_BP        1.278e-03  7.531e-04   1.697  0.08971 .
## glucose_conc:Triceps_thk         1.857e-03  1.086e-03   1.710  0.08730 .
## glucose_conc:X2_hr_insulin      -4.682e-05  6.523e-05  -0.718  0.47296
## glucose_conc:BMI                -4.981e-03  1.838e-03  -2.710  0.00672 **
## glucose_conc:Pedigree            5.903e-03  2.858e-02   0.207  0.83638
## glucose_conc:Age                -7.036e-04  1.111e-03  -0.633  0.52642
## glucose_conc:category_pregnant1 -6.840e-03  3.487e-02  -0.196  0.84449
## glucose_conc:category_pregnant2 -5.609e-02  3.330e-02  -1.684  0.09212 .
## glucose_conc:category_pregnant3 -1.518e-02  2.908e-02  -0.522  0.60164
## glucose_conc:category_pregnant4 -4.070e-02  3.252e-02  -1.252  0.21063
## Diastolic_BP:Triceps_thk        -3.189e-04  2.138e-03  -0.149  0.88145
## Diastolic_BP:X2_hr_insulin      -4.101e-04  2.017e-04  -2.034  0.04200 *
## Diastolic_BP:BMI                 1.723e-03  2.503e-03   0.688  0.49123
## Diastolic_BP:Pedigree           -3.592e-02  5.549e-02  -0.647  0.51745
```

```
## Diastolic_BP:Age                  4.995e-04  2.488e-03   0.201  0.84087
## Diastolic_BP:category_pregnant1   3.205e-02  5.462e-02   0.587  0.55740
## Diastolic_BP:category_pregnant2   1.131e-01  6.581e-02   1.719  0.08561 .
## Diastolic_BP:category_pregnant3   2.896e-02  5.939e-02   0.488  0.62576
## Diastolic_BP:category_pregnant4   6.343e-02  7.248e-02   0.875  0.38149
## Triceps_thk:X2_hr_insulin         2.088e-05  2.730e-04   0.076  0.93903
## Triceps_thk:BMI                  -3.840e-03  3.235e-03  -1.187  0.23513
## Triceps_thk:Pedigree              3.577e-02  8.818e-02   0.406  0.68503
## Triceps_thk:Age                  -7.327e-03  3.758e-03  -1.949  0.05124 .
## Triceps_thk:category_pregnant1   -6.943e-02  9.128e-02  -0.761  0.44685
## Triceps_thk:category_pregnant2   -5.485e-03  8.451e-02  -0.065  0.94825
## Triceps_thk:category_pregnant3    8.371e-02  7.762e-02   1.079  0.28081
## Triceps_thk:category_pregnant4    4.781e-02  9.416e-02   0.508  0.61160
## X2_hr_insulin:BMI                 4.947e-04  4.405e-04   1.123  0.26144
## X2_hr_insulin:Pedigree           -1.403e-02  5.990e-03  -2.342  0.01919 *
## X2_hr_insulin:Age                 6.816e-04  3.335e-04   2.044  0.04098 *
## X2_hr_insulin:category_pregnant1  2.805e-03  7.817e-03   0.359  0.71972
## X2_hr_insulin:category_pregnant2  7.875e-03  7.958e-03   0.990  0.32235
## X2_hr_insulin:category_pregnant3  2.324e-03  6.652e-03   0.349  0.72686
## X2_hr_insulin:category_pregnant4  1.473e-03  8.583e-03   0.172  0.86375
## BMI:Pedigree                     -1.636e-02  1.352e-01  -0.121  0.90366
## BMI:Age                           9.084e-03  5.553e-03   1.636  0.10190
## BMI:category_pregnant1            1.307e-01  1.476e-01   0.886  0.37580
## BMI:category_pregnant2           -3.460e-02  1.255e-01  -0.276  0.78285
## BMI:category_pregnant3           -2.284e-01  1.309e-01  -1.745  0.08103 .
## BMI:category_pregnant4           -1.310e-01  1.421e-01  -0.922  0.35650
## Pedigree:Age                     -7.049e-02  8.266e-02  -0.853  0.39381
## Pedigree:category_pregnant1      -7.050e-01  2.184e+00  -0.323  0.74681
## Pedigree:category_pregnant2      -1.223e+00  2.065e+00  -0.592  0.55357
## Pedigree:category_pregnant3       1.513e+00  2.160e+00   0.701  0.48353
## Pedigree:category_pregnant4       1.051e-01  2.423e+00   0.043  0.96539
## Age:category_pregnant1            3.400e-01  1.277e-01   2.662  0.00777 **
## Age:category_pregnant2            3.001e-01  1.822e-01   1.647  0.09950 .
## Age:category_pregnant3            1.612e-01  1.123e-01   1.436  0.15103
## Age:category_pregnant4            1.743e-01  1.138e-01   1.531  0.12582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 269.17  on 331  degrees of freedom
## AIC: 391.17
##
## Number of Fisher Scoring iterations: 7
```

Since the model size is too large and many non-significant variables are included, we performed LRT test for the interaction terms. From the test, 5 interaction terms are significant. 1) glucose:BMI, 2) thickness:age, 3) insulin:pedigree, 4) insulin:age, and 5) Age:pregnancies.

```
#Perform LRT test for the interaction terms
drop1(model5, test="LRT")
```

```
## Single term deletions
##
```

```
## Model:
## Diabetes ~ (glucose_conc + Diastolic_BP + Triceps_thk + X2_hr_insulin +
##     BMI + Pedigree + Age + category_pregnant)^2
##                                 Df Deviance    AIC    LRT Pr(>Chi)
## <none>                               269.17 391.17
## glucose_conc:Diastolic_BP        1   272.19 392.19 3.0162 0.082435 .
## glucose_conc:Triceps_thk         1   272.24 392.24 3.0681 0.079845 .
## glucose_conc:X2_hr_insulin       1   269.69 389.69 0.5133 0.473710
## glucose_conc:BMI                 1   276.88 396.88 7.7068 0.005501 **
## glucose_conc:Pedigree            1   269.21 389.21 0.0429 0.835988
## glucose_conc:Age                 1   269.57 389.57 0.3953 0.529546
## glucose_conc:category_pregnant   4   274.00 388.00 4.8241 0.305824
## Diastolic_BP:Triceps_thk         1   269.19 389.19 0.0222 0.881560
## Diastolic_BP:X2_hr_insulin       1   273.61 393.61 4.4395 0.035117 *
## Diastolic_BP:BMI                 1   269.64 389.64 0.4635 0.495994
## Diastolic_BP:Pedigree            1   269.62 389.62 0.4434 0.505491
## Diastolic_BP:Age                 1   269.21 389.21 0.0403 0.840878
## Diastolic_BP:category_pregnant   4   273.00 387.00 3.8286 0.429695
## Triceps_thk:X2_hr_insulin        1   269.18 389.18 0.0059 0.938876
## Triceps_thk:BMI                  1   270.62 390.62 1.4492 0.228651
## Triceps_thk:Pedigree             1   269.34 389.34 0.1664 0.683346
## Triceps_thk:Age                  1   273.27 393.27 4.1000 0.042882 *
## Triceps_thk:category_pregnant    4   273.69 387.69 4.5209 0.340071
## X2_hr_insulin:BMI                1   270.31 390.31 1.1368 0.286340
## X2_hr_insulin:Pedigree           1   276.00 396.00 6.8271 0.008978 **
## X2_hr_insulin:Age                1   273.90 393.90 4.7243 0.029740 *
## X2_hr_insulin:category_pregnant  4   270.37 384.37 1.1969 0.878608
## BMI:Pedigree                     1   269.19 389.19 0.0146 0.903731
## BMI:Age                          1   271.95 391.95 2.7801 0.095441 .
## BMI:category_pregnant            4   277.56 391.56 8.3835 0.078497 .
## Pedigree:Age                     1   269.90 389.90 0.7287 0.393302
## Pedigree:category_pregnant       4   271.55 385.55 2.3751 0.667138
## Age:category_pregnant            4   279.07 393.07 9.8962 0.042213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The five variables are added to the original model. We performed the Wald test for the coefficient again. In this time, the most non-significant variable is removed one by one. From the result, BP was the least significant with p-value, 0.7031, and removed.

```
#Fit the model which has 8 single variables and 5 interaction terms
model6 <- glm(Diabetes ~ glucose_conc+Diastolic_BP+Triceps_thk+X2_hr_insulin+BMI+Pedigree+Age+category_
#Check p-value for Wald test
summary(model6)
```

```
##
## Call:
## glm(formula = Diabetes ~ glucose_conc + Diastolic_BP + Triceps_thk +
##     X2_hr_insulin + BMI + Pedigree + Age + category_pregnant +
##     glucose_conc:BMI + Triceps_thk:Age + X2_hr_insulin:Pedigree +
##     X2_hr_insulin:Age + Age:category_pregnant, family = binomial,
##     data = db)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)            -1.406e+01  4.885e+00  -2.879 0.003989 **
## glucose_conc            8.437e-02  3.123e-02   2.702 0.006897 **
## Diastolic_BP           -4.310e-03  1.296e-02  -0.332 0.739529
## Triceps_thk             8.608e-02  5.331e-02   1.615 0.106409
## X2_hr_insulin          -6.786e-03  4.869e-03  -1.394 0.163394
## BMI                     2.206e-01  1.148e-01   1.922 0.054599 .
## Pedigree                2.982e+00  7.113e-01   4.192 2.76e-05 ***
## Age                    -7.299e-02  8.054e-02  -0.906 0.364809
## category_pregnant1     -7.551e+00  2.131e+00  -3.544 0.000394 ***
## category_pregnant2     -4.865e+00  3.568e+00  -1.364 0.172649
## category_pregnant3     -2.493e+00  1.862e+00  -1.339 0.180709
## category_pregnant4     -1.249e+00  2.032e+00  -0.614 0.538899
## glucose_conc:BMI       -1.193e-03  8.876e-04  -1.344 0.178974
## Triceps_thk:Age        -2.181e-03  1.561e-03  -1.397 0.162326
## X2_hr_insulin:Pedigree -6.931e-03  2.259e-03  -3.068 0.002153 **
## X2_hr_insulin:Age       3.204e-04  1.424e-04   2.250 0.024447 *
## Age:category_pregnant1  2.664e-01  7.576e-02   3.517 0.000437 ***
## Age:category_pregnant2  1.836e-01  1.355e-01   1.354 0.175583
## Age:category_pregnant3  1.133e-01  6.685e-02   1.695 0.090157 .
## Age:category_pregnant4  8.600e-02  6.461e-02   1.331 0.183176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 308.42  on 372  degrees of freedom
## AIC: 348.42
##
## Number of Fisher Scoring iterations: 6
```

Using the model from the previous test, we performed the test again. In this test, Age variable has the largest p-value, 0.364. The Age varialbe is removed in this stage.

```
#Fit the model from the previous test
model7 <- glm(Diabetes ~ glucose_conc+Triceps_thk+X2_hr_insulin+BMI+Pedigree+Age+category_pregnant+gluc
#Check p-value for Wald test
summary(model7)
```

```
##
## Call:
## glm(formula = Diabetes ~ glucose_conc + Triceps_thk + X2_hr_insulin +
##     BMI + Pedigree + Age + category_pregnant + glucose_conc:BMI +
##     Triceps_thk:Age + X2_hr_insulin:Pedigree + X2_hr_insulin:Age +
##     Age:category_pregnant, family = binomial, data = db)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.434e+01  4.815e+00  -2.978 0.002900 **
## glucose_conc            8.522e-02  3.114e-02   2.737 0.006202 **
## Triceps_thk             8.487e-02  5.332e-02   1.592 0.111445
## X2_hr_insulin          -6.810e-03  4.881e-03  -1.395 0.162950
## BMI                     2.227e-01  1.148e-01   1.940 0.052383 .
## Pedigree                2.987e+00  7.110e-01   4.201 2.66e-05 ***
## Age                    -7.547e-02  8.026e-02  -0.940 0.347052
```

```
## category_pregnant1     -7.470e+00  2.113e+00  -3.535 0.000407 ***
## category_pregnant2     -4.857e+00  3.567e+00  -1.361 0.173398
## category_pregnant3     -2.514e+00  1.862e+00  -1.350 0.177016
## category_pregnant4     -1.222e+00  2.033e+00  -0.601 0.547735
## glucose_conc:BMI       -1.225e-03  8.825e-04  -1.388 0.165157
## Triceps_thk:Age        -2.153e-03  1.565e-03  -1.376 0.168919
## X2_hr_insulin:Pedigree -6.920e-03  2.261e-03  -3.060 0.002215 **
## X2_hr_insulin:Age       3.220e-04  1.430e-04   2.252 0.024308 *
## Age:category_pregnant1  2.642e-01  7.538e-02   3.505 0.000457 ***
## Age:category_pregnant2  1.839e-01  1.355e-01   1.357 0.174687
## Age:category_pregnant3  1.140e-01  6.683e-02   1.706 0.087933 .
## Age:category_pregnant4  8.558e-02  6.465e-02   1.324 0.185581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 308.53  on 373  degrees of freedom
## AIC: 346.53
##
## Number of Fisher Scoring iterations: 6
```

We keep performing the same test and removing the most non-significant variables until only significant variables are left. At the end, the model 4 is obtained.

```
#Fit the model get from the previous test
model8 <- glm(Diabetes ~ glucose_conc+Triceps_thk+X2_hr_insulin+BMI+Pedigree+category_pregnant+glucose_
#Check p-value for Wald test
summary(model8)
```

```
##
## Call:
## glm(formula = Diabetes ~ glucose_conc + Triceps_thk + X2_hr_insulin +
##     BMI + Pedigree + category_pregnant + glucose_conc:BMI + Triceps_thk:Age +
##     X2_hr_insulin:Pedigree + X2_hr_insulin:Age + Age:category_pregnant,
##     family = binomial, data = db)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.671e+01  4.145e+00  -4.031 5.55e-05 ***
## glucose_conc            8.801e-02  3.103e-02   2.837 0.004560 **
## Triceps_thk             1.141e-01  4.428e-02   2.577 0.009977 **
## X2_hr_insulin          -5.015e-03  4.411e-03  -1.137 0.255499
## BMI                     2.322e-01  1.143e-01   2.031 0.042253 *
## Pedigree                2.944e+00  7.129e-01   4.129 3.64e-05 ***
## category_pregnant1     -6.543e+00  1.881e+00  -3.479 0.000504 ***
## category_pregnant2     -3.696e+00  3.345e+00  -1.105 0.269183
## category_pregnant3     -1.573e+00  1.585e+00  -0.993 0.320951
## category_pregnant4     -2.079e-01  1.720e+00  -0.121 0.903798
## glucose_conc:BMI       -1.334e-03  8.738e-04  -1.526 0.126935
## Triceps_thk:Age        -3.049e-03  1.268e-03  -2.404 0.016226 *
## X2_hr_insulin:Pedigree -6.890e-03  2.287e-03  -3.013 0.002586 **
## X2_hr_insulin:Age       2.675e-04  1.259e-04   2.124 0.033677 *
## category_pregnant1:Age  2.265e-01  6.407e-02   3.536 0.000406 ***
```

```
## category_pregnant2:Age  1.387e-01  1.264e-01   1.097 0.272469
## category_pregnant3:Age  7.650e-02  5.399e-02   1.417 0.156494
## category_pregnant4:Age  4.633e-02  4.957e-02   0.935 0.349940
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 309.43  on 374  degrees of freedom
## AIC: 345.43
##
## Number of Fisher Scoring iterations: 6
```

```
#Fit the model get from the previous test
model9 <- glm(Diabetes ~ glucose_conc+Triceps_thk+X2_hr_insulin+BMI+Pedigree+glucose_conc:BMI+Triceps_t
#Check p-value for Wald test
summary(model9)
```

```
##
## Call:
## glm(formula = Diabetes ~ glucose_conc + Triceps_thk + X2_hr_insulin +
##     BMI + Pedigree + glucose_conc:BMI + Triceps_thk:Age + X2_hr_insulin:Pedigree +
##     X2_hr_insulin:Age + Age:category_pregnant, family = binomial,
##     data = db)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.561e+01  3.951e+00  -3.951 7.80e-05 ***
## glucose_conc            8.081e-02  2.990e-02   2.703  0.00688 **
## Triceps_thk             5.270e-02  3.465e-02   1.521  0.12829
## X2_hr_insulin          -7.036e-03  4.299e-03  -1.637  0.10170
## BMI                     2.075e-01  1.100e-01   1.886  0.05927 .
## Pedigree                2.747e+00  6.795e-01   4.043 5.28e-05 ***
## glucose_conc:BMI       -1.103e-03  8.490e-04  -1.300  0.19377
## Triceps_thk:Age        -1.084e-03  9.180e-04  -1.180  0.23786
## X2_hr_insulin:Pedigree -6.769e-03  2.296e-03  -2.948  0.00320 **
## X2_hr_insulin:Age       3.226e-04  1.237e-04   2.608  0.00910 **
## Age:category_pregnant1  9.718e-03  1.799e-02   0.540  0.58908
## Age:category_pregnant2  4.345e-03  2.072e-02   0.210  0.83396
## Age:category_pregnant3  1.988e-02  1.747e-02   1.138  0.25522
## Age:category_pregnant4  1.571e-02  1.820e-02   0.863  0.38810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.1  on 391  degrees of freedom
## Residual deviance: 325.6  on 378  degrees of freedom
## AIC: 353.6
##
## Number of Fisher Scoring iterations: 5
```

```
#Fit the model get from the previous test
model10 <- glm(Diabetes ~ glucose_conc+Triceps_thk+X2_hr_insulin+BMI+Pedigree+glucose_conc:BMI+Triceps_
```

```r
#Check p-value for Wald test
summary(model10)
```

```
##
## Call:
## glm(formula = Diabetes ~ glucose_conc + Triceps_thk + X2_hr_insulin +
##     BMI + Pedigree + glucose_conc:BMI + Triceps_thk:Age + X2_hr_insulin:Pedigree +
##     X2_hr_insulin:Age, family = binomial, data = db)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.528e+01  3.924e+00  -3.893 9.88e-05 ***
## glucose_conc            8.318e-02  2.973e-02   2.797  0.00515 **
## Triceps_thk             3.700e-02  3.056e-02   1.211  0.22607
## X2_hr_insulin          -8.086e-03  4.235e-03  -1.909  0.05621 .
## BMI                     2.084e-01  1.097e-01   1.900  0.05741 .
## Pedigree                2.650e+00  6.675e-01   3.970 7.20e-05 ***
## glucose_conc:BMI       -1.155e-03  8.446e-04  -1.368  0.17136
## Triceps_thk:Age        -6.037e-04  7.347e-04  -0.822  0.41125
## X2_hr_insulin:Pedigree -6.422e-03  2.228e-03  -2.882  0.00395 **
## X2_hr_insulin:Age       3.539e-04  1.221e-04   2.897  0.00377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 327.29  on 382  degrees of freedom
## AIC: 347.29
##
## Number of Fisher Scoring iterations: 5
```

```r
#Fit the model get from the previous test
model11 <- glm(Diabetes ~ glucose_conc+Triceps_thk+X2_hr_insulin+BMI+Pedigree+glucose_conc:BMI+X2_hr_ins
#Check p-value for Wald test
summary(model11)
```

```
##
## Call:
## glm(formula = Diabetes ~ glucose_conc + Triceps_thk + X2_hr_insulin +
##     BMI + Pedigree + glucose_conc:BMI + X2_hr_insulin:Pedigree +
##     X2_hr_insulin:Age, family = binomial, data = db)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.521e+01  3.916e+00  -3.884 0.000103 ***
## glucose_conc            8.253e-02  2.962e-02   2.786 0.005335 **
## Triceps_thk             1.622e-02  1.753e-02   0.925 0.354747
## X2_hr_insulin          -5.666e-03  2.947e-03  -1.923 0.054532 .
## BMI                     2.116e-01  1.092e-01   1.938 0.052644 .
## Pedigree                2.633e+00  6.686e-01   3.938 8.21e-05 ***
## glucose_conc:BMI       -1.164e-03  8.397e-04  -1.386 0.165691
## X2_hr_insulin:Pedigree -6.363e-03  2.225e-03  -2.859 0.004246 **
## X2_hr_insulin:Age       2.765e-04  7.085e-05   3.902 9.54e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.1  on 391  degrees of freedom
## Residual deviance: 328.0  on 383  degrees of freedom
## AIC: 346
##
## Number of Fisher Scoring iterations: 5
```

```
#Fit the model get from the previous test
model12 <- glm(Diabetes ~ glucose_conc+X2_hr_insulin+BMI+Pedigree+glucose_conc:BMI+X2_hr_insulin:Pedigr
#Check p-value for Wald test
summary(model12)
```

```
##
## Call:
## glm(formula = Diabetes ~ glucose_conc + X2_hr_insulin + BMI +
##     Pedigree + glucose_conc:BMI + X2_hr_insulin:Pedigree + X2_hr_insulin:Age,
##     family = binomial, data = db)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.532e+01  3.902e+00  -3.925 8.66e-05 ***
## glucose_conc            8.316e-02  2.956e-02   2.813  0.00490 **
## X2_hr_insulin          -5.983e-03  2.930e-03  -2.042  0.04114 *
## BMI                     2.288e-01  1.074e-01   2.131  0.03311 *
## Pedigree                2.646e+00  6.641e-01   3.985 6.75e-05 ***
## glucose_conc:BMI       -1.178e-03  8.382e-04  -1.406  0.15982
## X2_hr_insulin:Pedigree -6.234e-03  2.185e-03  -2.853  0.00432 **
## X2_hr_insulin:Age       2.835e-04  7.124e-05   3.979 6.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 328.85  on 384  degrees of freedom
## AIC: 344.85
##
## Number of Fisher Scoring iterations: 5
```

```
#Fit the model get from the previous test
model4 <- glm(Diabetes ~ glucose_conc+X2_hr_insulin+BMI+Pedigree+X2_hr_insulin:Pedigree+X2_hr_insulin:A
#Check p-value for Wald test
summary(model4)
```

```
##
## Call:
## glm(formula = Diabetes ~ glucose_conc + X2_hr_insulin + BMI +
##     Pedigree + X2_hr_insulin:Pedigree + X2_hr_insulin:Age, family = binomial,
##     data = db)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.020e+01  1.167e+00  -8.739  < 2e-16 ***
```

```
## glucose_conc            4.306e-02  6.072e-03   7.091 1.33e-12 ***
## X2_hr_insulin          -6.848e-03  2.907e-03  -2.356 0.018480 *
## BMI                     8.193e-02  2.163e-02   3.787 0.000152 ***
## Pedigree                2.590e+00  6.585e-01   3.933 8.39e-05 ***
## X2_hr_insulin:Pedigree -6.132e-03  2.185e-03  -2.807 0.005005 **
## X2_hr_insulin:Age       2.939e-04  7.098e-05   4.141 3.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 330.83  on 385  degrees of freedom
## AIC: 344.83
##
## Number of Fisher Scoring iterations: 5
```

We performed another method to find the model. In this time we added the interaction terms. First we checked the deviance of null model and original model. the residual deviance is decreased significantly with the inclusion of predictor variables, it can be said that the model with the inclusion of predictor variables is better than the null model. Then we test the original model's coefficient what we did before. Model 3 is constructed using the variables which are significant in model2. Using model 3, we fitted seven models. Each model has four single variable s and interaction terms of each 8 single variable. (e.g Diabetes ~ BMI + glucose + pedigree + insulin + BMI's all interaction terms) From these models, significant interaction terms are found and these terms are added to model 3. We tested the model to check whether the added terms are significant or not. From that, we found that insulin with pedigree and age is significant but the single age is not. Therefore age is removed. Since insulin variable has two significant interaction terms so we added the insulin and tested. The result is that insulin is significant, so we keep the insulin. Finally, we obtained the model which is same as cutting down method.

```
model1 <- glm(Diabetes ~ 1 , data = db, family = binomial)
fit1 <- summary(model1)
#null model

model2 <- glm(Diabetes ~ . - times_pregnant , data = db, family = binomial)   ## fit the model
#summary(cat.glm) #summary of fit
fit2 <- summary(model2)

model3 <- glm(Diabetes ~ BMI + glucose_conc + Pedigree + Age , data = db, family = binomial)
#model removed non-significant variables
fit3 <- summary(model3)

model4 <- glm(Diabetes ~ BMI + glucose_conc + Pedigree + X2_hr_insulin + Pedigree:X2_hr_insulin + Age:X
fit4 <- summary(model4) #331

model5 <- glm(Diabetes ~ glucose_conc+X2_hr_insulin+BMI+Pedigree+glucose_conc:BMI+X2_hr_insulin:Pedigre
fit5 <- summary(model5)
```

10.3 Confidence Interval Estimation

We could also check the hypothesis test result by calculating confidence interval. First we calculated the original model's confidence interval. Same as the hypothesis test, only 4 variables' CI does not include 0 values. In other words, the confidence intervals which does not include zero are significant. The significant variables are glucose, BMI, pedigree, and age.

```r
#Coefficient estimation of original model
ori_intercept <- summary(model2)$coefficients["(Intercept)","Estimate"]
ori_glucose_conc <- summary(model2)$coefficients["glucose_conc","Estimate"]
ori_Diastolic_BP <- summary(model2)$coefficients["Diastolic_BP","Estimate"]
ori_Triceps_thk <- summary(model2)$coefficients["Triceps_thk","Estimate"]
ori_X2_hr_insulin <-  summary(model2)$coefficients["X2_hr_insulin","Estimate"]
ori_BMI <-  summary(model2)$coefficients["BMI","Estimate"]
ori_Pedigree <-  summary(model2)$coefficients["Pedigree","Estimate"]
ori_Age <-  summary(model2)$coefficients["Age","Estimate"]
ori_pregnant1 <-  summary(model2)$coefficients["category_pregnant1","Estimate"]
ori_pregnant2 <-  summary(model2)$coefficients["category_pregnant2","Estimate"]
ori_pregnant3 <-  summary(model2)$coefficients["category_pregnant3","Estimate"]
ori_pregnant4 <-  summary(model2)$coefficients["category_pregnant4","Estimate"]

#Standard error of original model coefficient
ori_intercept_se <- summary(model2)$coefficients["(Intercept)","Std. Error"]
ori_glucose_conc_se <- summary(model2)$coefficients["glucose_conc","Std. Error"]
ori_Diastolic_BP_se <- summary(model2)$coefficients["Diastolic_BP","Std. Error"]
ori_Triceps_thk_se <- summary(model2)$coefficients["Triceps_thk","Std. Error"]
ori_X2_hr_insulin_se <-  summary(model2)$coefficients["X2_hr_insulin","Std. Error"]
ori_BMI_se <-  summary(model2)$coefficients["BMI","Std. Error"]
ori_Pedigree_se <-  summary(model2)$coefficients["Pedigree","Std. Error"]
ori_Age_se <-  summary(model2)$coefficients["Age","Std. Error"]
ori_pregnant1_se <-  summary(model2)$coefficients["category_pregnant1","Std. Error"]
ori_pregnant2_se <-  summary(model2)$coefficients["category_pregnant2","Std. Error"]
ori_pregnant3_se <-  summary(model2)$coefficients["category_pregnant3","Std. Error"]
ori_pregnant4_se <-  summary(model2)$coefficients["category_pregnant4","Std. Error"]

#Calculating 95% of CI of original model
ori_intercept_ci <- c(ori_intercept - 1.96 * ori_intercept_se, ori_intercept + 1.96 * ori_intercept_se)
ori_glucose_conc_ci <- c(ori_glucose_conc - 1.96 * ori_glucose_conc_se, ori_glucose_conc + 1.96 * ori_gl
ori_Diastolic_BP_ci <- c(ori_Diastolic_BP - 1.96 * ori_Diastolic_BP_se, ori_Diastolic_BP + 1.96 * ori_Di
ori_Triceps_thk_ci <- c(ori_Triceps_thk - 1.96 * ori_Triceps_thk_se, ori_Triceps_thk + 1.96 * ori_Tricep
ori_X2_hr_insulin_ci <-  c(ori_X2_hr_insulin - 1.96 * ori_X2_hr_insulin_se, ori_X2_hr_insulin + 1.96 * o
ori_BMI_ci <-  c(ori_BMI - 1.96 * ori_BMI_se, ori_BMI + 1.96 * ori_BMI_se)
ori_Pedigree_ci <-  c(ori_Pedigree - 1.96 * ori_Pedigree_se, ori_Pedigree + 1.96 * ori_Pedigree_se)
ori_Age_ci <-  c(ori_Age - 1.96 * ori_Age_se, ori_Age + 1.96 * ori_Age_se)
ori_pregnant1_ci <-  c(ori_pregnant1 - 1.96 * ori_pregnant1_se, ori_pregnant1 + 1.96 * ori_pregnant1_se)
ori_pregnant2_ci <-  c(ori_pregnant2 - 1.96 * ori_pregnant2_se, ori_pregnant2 + 1.96 * ori_pregnant2_se)
ori_pregnant3_ci <-  c(ori_pregnant3_se - 1.96 * ori_pregnant3_se, ori_pregnant3_se + 1.96 * ori_pregnan
ori_pregnant4_ci <-  c(ori_pregnant4_se - 1.96 * ori_pregnant4_se, ori_pregnant4_se + 1.96 * ori_pregnan

#Confidence interval of original model
ori_intercept_ci
ori_glucose_conc_ci
ori_Diastolic_BP_ci
ori_Triceps_thk_ci
ori_X2_hr_insulin_ci
ori_BMI_ci
ori_Pedigree_ci
ori_Age_ci
ori_pregnant1_ci
ori_pregnant2_ci
```

```
ori_pregnant3_ci
ori_pregnant4_ci
```

We also checked the confidence interval of final models. The final model has only the significant variables, so zero is not included in each confidence interval.

```
#Coefficient estimation of final model
intercept <- summary(model4)$coefficients["(Intercept)","Estimate"]
glucose_conc <- summary(model4)$coefficients["glucose_conc","Estimate"]
X2_hr_insulin <-  summary(model4)$coefficients["X2_hr_insulin","Estimate"]
BMI <-  summary(model4)$coefficients["BMI","Estimate"]
Pedigree <-  summary(model4)$coefficients["Pedigree","Estimate"]
X2_hr_insulin_Pedigree <-  summary(model4)$coefficients["Pedigree:X2_hr_insulin","Estimate"]
X2_hr_insulin_Age <-  summary(model4)$coefficients["X2_hr_insulin:Age","Estimate"]

#Standard error of final model coefficient
intercept_se <- summary(model4)$coefficients["(Intercept)","Std. Error"]
glucose_conc_se <- summary(model4)$coefficients["glucose_conc","Std. Error"]
X2_hr_insulin_se <-  summary(model4)$coefficients["X2_hr_insulin","Std. Error"]
BMI_se <-  summary(model4)$coefficients["BMI","Std. Error"]
Pedigree_se <-  summary(model4)$coefficients["Pedigree","Std. Error"]
X2_hr_insulin_Pedigree_se <-  summary(model4)$coefficients["Pedigree:X2_hr_insulin","Std. Error"]
X2_hr_insulin_Age_se <-  summary(model4)$coefficients["X2_hr_insulin:Age","Std. Error"]

#Calculating 95% of CI of original model
intercept_ci <- c(intercept - 1.96 * intercept_se, intercept + 1.96 * intercept_se)
glucose_conc_ci <- c(glucose_conc - 1.96 * glucose_conc_se, glucose_conc + 1.96 * glucose_conc_se)
X2_hr_insulin_ci <-  c(X2_hr_insulin - 1.96 * X2_hr_insulin_se, X2_hr_insulin + 1.96 * X2_hr_insulin_se)
BMI_ci <-  c(BMI - 1.96 * BMI_se, BMI + 1.96 * BMI_se)
Pedigree_ci <-  c(Pedigree - 1.96 * Pedigree_se, Pedigree + 1.96 * Pedigree_se)
X2_hr_insulin_Pedigree_ci <-  c(X2_hr_insulin_Pedigree - 1.96 * X2_hr_insulin_Pedigree_se, X2_hr_insulin
X2_hr_insulin_Age_ci <-  c(X2_hr_insulin_Age - 1.96 * X2_hr_insulin_Age_se, X2_hr_insulin_Age + 1.96 *

#Confidence interval of final model
intercept_ci
```

```
## [1] -12.490724  -7.914178
```

```
glucose_conc_ci
```

```
## [1] 0.03115557 0.05495748
```

```
X2_hr_insulin_ci
```

```
## [1] -0.012546168 -0.001150777
```

```
BMI_ci
```

```
## [1] 0.03953247 0.12433598
```

```
Pedigree_ci
```

```
## [1] 1.299200 3.880433
```

```
X2_hr_insulin_Pedigree_ci
```

```
## [1] -0.010414135 -0.001849852
```

```
X2_hr_insulin_Age_ci
```

```
## [1] 0.0001548208 0.0004330618
```

10.4 Goodness-of-fit test

1. McFadden's Pseudo R-squared Test

We obtained the R^2 of 4 candidate models.

```
r1 <- (model1$null.deviance-model1$deviance)/model1$null.deviance
r2 <-(model2$null.deviance-model2$deviance)/model2$null.deviance
r3 <-(model3$null.deviance-model3$deviance)/model3$null.deviance
r4 <-(model4$null.deviance-model4$deviance)/model4$null.deviance
cat(r1,r2,r3,r4)
```

```
## 1.14121e-16 0.3073634 0.3028779 0.3358161
```

2. Deviance Test

Residual deviance and corresponding p-value for 4 models.

```
##aic
cat(fit1$aic,fit2$aic,fit3$aic,fit4$aic,fit5$aic)
```

```
## 500.0978 369.0008 357.235 344.8285 344.8527
```

```
#chi1 <- sum(residuals(model1, type = "pearson")^2)
deviance1 <- deviance(model1)
deviance2 <- deviance(model2)
deviance3 <- deviance(model3)
deviance4 <- deviance(model4)
cat("Deviance",deviance1,deviance2,deviance3,deviance4)
```

```
## Deviance 498.0978 345.0008 347.235 330.8285
```

```
p1 <- pchisq(fit1$deviance,fit1$df.residual, lower.tail = FALSE)
p2 <- pchisq(fit2$deviance,fit2$df.residual, lower.tail = FALSE)
p3 <- pchisq(fit3$deviance,fit3$df.residual, lower.tail = FALSE)
p4 <- pchisq(fit4$deviance,fit4$df.residual, lower.tail = FALSE)
cat("p-value",p1,p2,p3,p4)
```

```
## p-value 0.0001922302 0.9008472 0.9274781 0.9787088
```

10.5 Model Comparison Using AIC and BIC

We obtained the AIC, relative likelihood $\exp((\text{AIC}_{min} - \text{AIC}_i)/2)$ and BIC values using R function.

```
cat(fit1$aic,fit2$aic,fit3$aic,fit4$aic)#AIC
```

```
## 500.0978 369.0008 357.235 344.8285
```

```
rl1 <- exp((fit4$aic-fit1$aic)/2)#relative likelihood
rl2 <- exp((fit4$aic-fit2$aic)/2)
rl3 <- exp((fit4$aic-fit3$aic)/2)
cat(rl1,rl2,rl3)
```

```
## 1.92178e-34 5.637182e-06 0.002022876
```

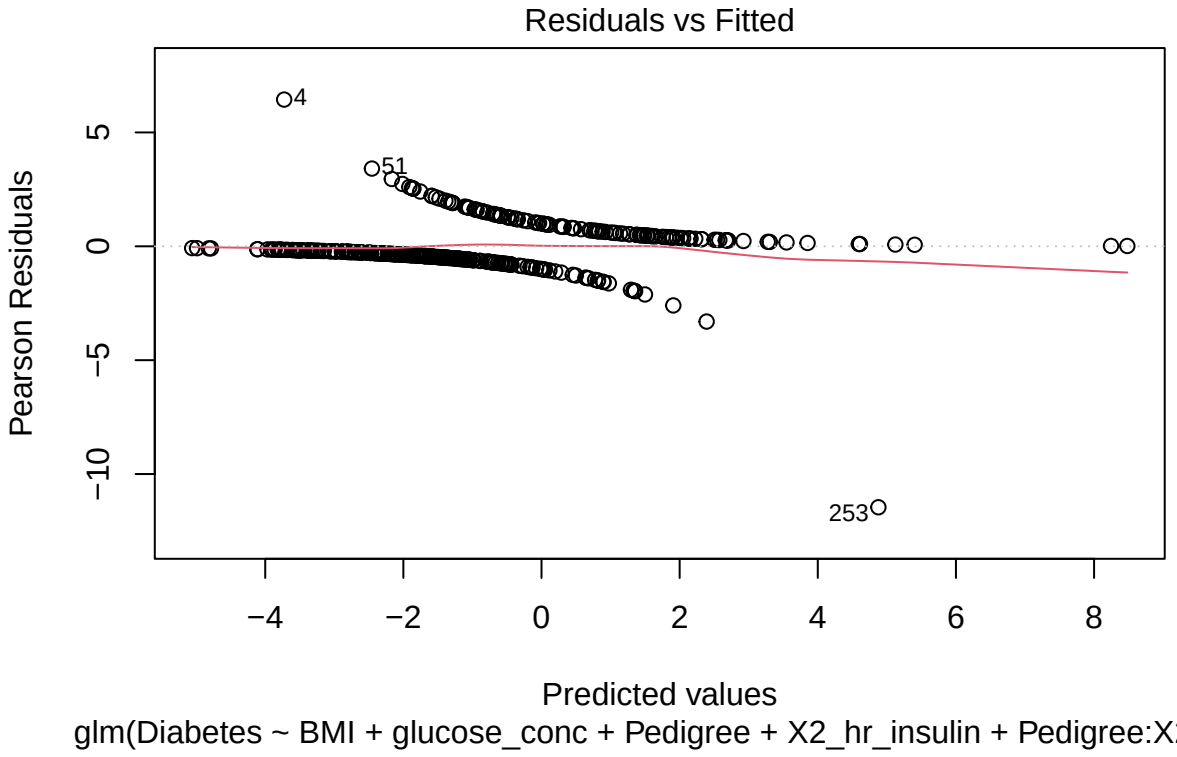
```
cat(BIC(model1),BIC(model2),BIC(model3),BIC(model4))
```

```
## 504.0691 416.6559 377.0913 372.6274
```

10.6 Plots of the final model

Residuals vs Fitted plot and Q-Q plot.

```
plot(model4, 1)
```




```
plot(model4, 2)
```