Logistic Regression for the Analysis of Diabetic Data

Smita Sukhadeve

December 16, 2016

1 Introduction

This paper focuses on Logistic Regression modelling for the analysis of diabetes data. Sample data used during the study obtained from the UCI- Machine Learning repository and constitutes information about 768 Pima Indian females older than 20 years. Originally, 'National Institute of Diabetes and Digestive and Kidney Disease' collected this data to analyse the cause for high incidence rate of diabetes in this population. The response 'outcome' is a binary and indicates whether the subject is diabetic or not(1/0). Logit models are suitable when the response is binary. Therefore, we apply logistic regression to research on following questions:

- What are the significant risk factors which accounts for the high diabetic rate among this population
- Comparative study based on the different population groups
- How various attributes are associated with diagnosis of diabetes
- Predictive power of logistic regression to correctly identify the diabetic and non-diabetic patient

2 Description of Data

There are total 8 explanatory variables in the data which describes health related attributes of the subjects. Additional information about these variables is as follows:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age in years

All the variables are continuous/discrete. For the purpose of this study, we will create three categorical variables based on BMI, Blood Pressures and Age. Based on the patient's age, we group them into 3 categories. Possible values are age35, age350, age> 50. We used wikipedia to categorise patients by their blood pressure and BMI values. Patient blood pressure ranges from low to high. We discretize bmi into underwt, normal, overwt and obess. Figure-1 shows proportion of subjects by these categorical variables. We have around 34% diabetic population in our sample. We observe the positive correlation between insulin

level with skin thickness and glucose level[Appendix-I]. Initial data exploration does not indicate presence of any missing values or outliers.

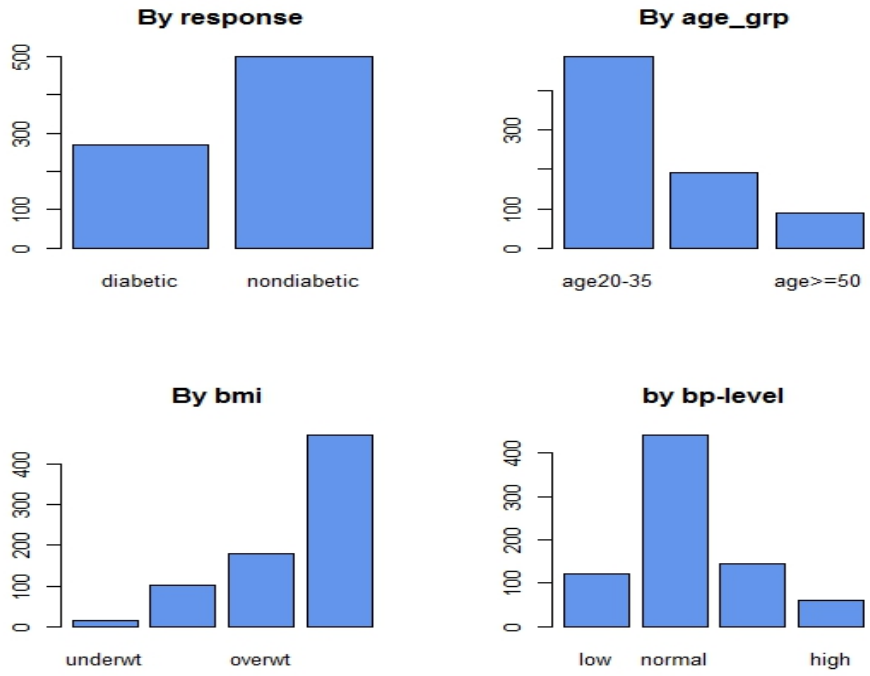


Figure 1: Frequency distribution by outcome, bmi, age, bp

3 Methods

3.1 Logistic Regression:

We consider logistic regression to build the model since our response is binary. This is one of the most popular generalized linear model for analysing the binary data. For Logit models, we assume binomial distribution of the response. Instead of modelling response directly, we model the probability of success. Definition of success changes with the problem. In our case, we define success as subject is diagnosed with diabetes and we model the probability that a subject is diagnosed with diabetes. Logistic regression model is given as follows:

$$logit(\pi(x)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p$$

The components of this generalized linear model are as follows:
Link function = $logit(\pi(x)) = log(\frac{\pi(x)}{1-\pi(x)})$
Systematic component = $\alpha + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p$

Random component = $\pi(x)$ which is success probability, i.e. $P(x = 1)$

We discuss in detail about this glm model in the technical section.

3.2 Variable Selection

Model with too many variables is often difficult to interpret. Moreover, there is always chance of over-fitting a model when we introduce extra variables. To avoid this, we will use variable selection approach. It helps to exclude the irrelevant features from the model. We use Akaike Information Criteria (AIC) to choose the optimal model. Model with a least AIC is considered as the best among multiple choices.
Akaike Information Criteria for a given model is calculated as below:

$$AIC = \frac{(RSS + 2d\hat{\sigma}^2)}{n\hat{\sigma}^2}$$

where,
n = number of observations
$\hat{\sigma}^2$ = Estimated variance of error associated with each response measurement
RSS = Residual sum of squares
d = Number of explanatory variables
Variable selection methods used in this step are as follows:

3.2.1 Stepwise Backward Selection

This selection method first consider the model with all the explanatory variables. It is iterative process where we drop variable at each step, refit the model and estimate 'Akaike Information Criteria'. For each iteration, we check AIC of model before and after removal of a predictor. We decide to keep removed predictor if the AIC of model increases after removal of this predictor. Finally, we choose the model with least value of AIC.

3.2.2 Best Subset Selection

This selection method is an extreme approach to find the best model as it considers all 2^p model choices. p represents number of explanatory variables we wish to consider for model building. Therefore, this technique is suitable only when we have small predictor set. We have total 8 explanatory variables. Moreover, size of dataset is relatively small. Hence, we choose the exhaustive search to find the optimal model. The process starts with null model with no variables and gradually fit all $\binom{8}{p}$ models and choose best among them using AIC metric.

Next, we introduce second order interaction terms in our full model and again apply variable selection method to find the best possible model using interaction terms.

3.3 Model Evaluation

We consider following tests to evaluate the performance of the selected models.

3.3.1 Global-Fit Statistics

We consider Pearson(χ^2) and Likelihood ratio(G^2) goodness-of-fit statistics to measure how well the selected model fits the data. The null hypothesis for both the global-Fit test is selected model fits the data. It summarizes how observed and fitted values differ from each other.

3.3.2 Residual analysis

Next, we analyse the residuals of these models by calculating Pearson and Deviance residual. These values tell us calculated difference between the observed and fitted values.

3.3.3 Lack-of-fit test

We use Hosmer-Lemeshow (HL) test statistic for performing the lack-of-fit test on the selected models. Null hypothesis for this test is given model fits the data. In Hosmer-Lemeshow test, we divide the data into different groups based on the estimated probabilities. We then take sum of all the square difference between the observed and fitted probabilities to get the HL test statistics. Small value of p represents given model doesn't fit the data well.

We analyse result of above mentioned tests and choose the best model among the two.

3.4 Model Inference

For the model inference, we follow below steps:

- Use Wald and Likelihood ratio test to evaluate the significance of the parameters. These are chi-square tests. Null hypothesis for the mentioned test statistics is $H_0 = \beta_0$.
- Analyse the confidence interval for the parameters. The 95% confidence interval for β is $[\beta \pm Z_{\frac{\alpha}{2}} * ASE]$
- Examine the confidence interval for odd Ratio and probability estimates. Logistic regression fitted probabilities are as calculated as below:

$$\pi(X = x) = \frac{exp(\hat{\alpha} + \hat{\beta}x)}{1 + exp(\hat{\alpha} + \hat{\beta}x)}$$

  We first compute the confidence interval for the logit(x).Then we use above formula and plugin value of x to find the confidence interval for estimated probabilities. We discuss more about the computing steps in the technical section.

- Analyse the distribution of fitted probabilities, residuals, marginal plots for the different groups.
- Evaluate model prediction power using area under the receiver operating characteristic(ROC) curve.

4 Result

4.1 Model-1 Using the Variable Selection Method

We have very few explanatory variables. Considering this, we applied both backward and subset selection methods.To fit the model, we have used 85% random sample to fit the model while remaining data was kept

for analysing the prediction power of the model. Fortunately, both the variable selection method found the same model which is given as below:

logit(π(Outcome))	= -6.70 + 0.113Pregnancies + 0.038Glucose - 0.0021Insulin + 1.05DiabetesPedigreeFunction -1.484bmiGrpnormal - 0.19bmiGrpoverwt + 0.94bmiGrpobess

where,
bmiGrpnormal= subjects under normal category
bmiGrpoverwt = subjects under overweight category
bmiGrpobess = subjects under obess category
$\pi(Outcome)$ =success probability that woman are diagnosed with diabetes

Residual deviance of the Null model was 832.31 on 651 degrees of freedom. This value decreased to 598.74 after including the explanatory variables with the loss of 7 degrees of freedom. We performed Pearson goodness-of-fit test to measure the model performance on the overall data. P-value using Chi-square test statistics for this model is 0.89 which is less than the significance level 0.05. Hence, we rejected the null hypothesis. Estimated model fitted the data well as compared to the null model. We performed the Deviance test using the R's anova() function to analyse the significant variables. Null model deviance dropped to 662.92 after the introduction of Glucose level. Moreover, deviance test statistics was found statistically significant(p-value < 0.05). Following table presents the result of deviance test and variables marked with '***' are significant at 0.05 significance level as (Chi square test statistics < 0.05).

Table-1 ANOVA result for model-1

Variable	Resid.Dev	Pr(>Chi)
NULL	832.30	NA
Pregnancies	806.16	$3.168e^{-7}$ ***
Glucose	662.92	$5.22e^{-33}$ ***
Insulin	661.81	0.29
DiabetesPedigreeFunction	648.71	$2.9e^{-4}$***
bmi_grp	598.73	$8.0e^{-11}$***

Following this step, we considered introducing the interaction terms in search of the good model.

4.2 Model With Interaction Term

We observed the variable selection methods did not consider the ageGrp during the selection process. Therefore, during this step, we included the variable ageGrp along with the optimal explanatory variables found during selection process to analyse the age effect on the occurrence of the diabetes. During the model search, we also introduced the second order interaction terms such as ageGrp*bmiGgrp, ageGrp*bpStatus, Glucose*Insulin to observe the interaction effect of these variables on the success probability. The backward selection was performed to find the optimal model with the interaction terms. We found following model using this approach:

logit(Outcome)	= Pregnancies + Glucose + Insulin + DiabetesPedigreeFunction + bmi_grp + age_grp + Insulin:ageGrp

Table-1 ANOVA result for model with Interaction term

Variable	Resid.Dev	Pr(>Chi)
NULL	832.30	NA
Pregnancies	806.16	$3.16e^{-7}$ ***
Glucose	662.92	$5.22e^{-33}$ ***
Insulin	661.81	0.29
DiabetesPedigreeFunction	648.71	2.94e-4$***$
bmiGrp	598.73	$8.08e^{-11}$***
ageGrp	596.89	0.39
Insulin:ageGrp	590.89	0.049*

We analysed the model 2 by Pearson goodness-of-fit test. P-value of (χ^2) test statistics is 0.91. This showed that our model fitted the data well since the p-value was greater than 0.05 significance. Next, we performed the deviance test to compare overall fitness of model 2. Result indicate that the residual deviance decreased from 832.3 to 590.8. Model 1 is nested within model 2. Therefore, we decided to compare their performance using the deviance test. We used anova() function to perform this test. We found the test statistic value is 0.09 which wasn't statistically significant. This shows that the additional interaction terms did not add significantly to the prediction of success probabilities. Hence, we could drop these additional interaction terms. However, for the scope this study we considered model 2 with interaction term as our final model.
In the next step, we measured the overall fitness of the model using lack of fit test with Hosmer and Lomeshow fit statistic. p-value for this test statistics is 0.061 which was still less the 0.05 significance level. Thus, both the fitness tests support our result that the model-2 did not fit the data well. However, we wanted to study the effect of age and Insulin level on the occurrence of diabetes. Therefore, we continued to use this model-2 as our final model.

4.3 Selection and interpretation of Final Model

Our Final Model is as follows:

logit(π(Outcome))	= -6.50+ 0.08Pregnancies + 0.038Glucose - 0.003Insulin + 1.04DiabetesPedigreeFunction -1.56bmiGrpnormal -0.31bmiGrpoverwt + 0.83bmiGrpobess + 0.19ageGrpage50 - 0.21ageGrpage> 50 + 0.002Insulin:ageGrpage50 + 0.005Insulin:ageGrpage> 50

The following table represents the estimated values of regression coefficient along with their standard errors.

Table 3.2 Regression Coefficient estimates for Model-3

Variable	Estm-Coeff	exp(Estm-Coeff)	Stad.Err.	p-value
(Intercept)	-6.50	0.0014	0.98	$3.44e^{-11}$ ***
Pregnancies	0.08	1.084	0.037	0.029 *
Glucose	0.038	1.039	0.004	$< 2e-16$ ***
Insulin	-0.003	0.997	0.001	0.00254 *
DiabetesPedigreeFunction	1.04	2.84	0.322	0.00120 **
bmi_grpnormal	-1.56	0.20	0.960	0.10
bmi_grpoverwt	-0.31	0.732	0.861	0.71
bmi_grpobess	0.83	2.30	0.843	0.32
age_grpage35-50	0.19	1.21	0.322	0.55
age_grpage$>= 50$	-0.21	0.80	0.419	0.60
Insulin:age_grpage35-50	0.002	1.002	0.002	0.31
Insulin:age_grpage$>= 50$	0.005	1.00	0.002	0.02 *

Note: Explanatory variables indicated with * found significant at some confidence level.

After analysing the Wald test statistics for the regression coefficients, we can see that Pregnancies, Glucose and Diabetes pedegree function found significant. The p-value for the Wald statistics of these variables is less that 0.05 significance level. Therefore, we rejected the null hypothesis for the Wald test. The response i.e. occurrence of Diabetes is dependent on these variables.

The value of regression coefficient for Glucose is 0.038 and exp(0.035)=1.039. This value implies that the log odds of the occurrence of diabetes increased by a factor 0.038 for 1 unit increment in glucose level by keeping all other variables fixed. In other word, the odd are multiplied by the factor 1.03 for a unit increase in the glucose level.

The regression coefficient for DiabetesPedigreeFunction is 1.04 and exp(1.04)= 2.84. This results shows log of odds of occurrence of diabetes increased by the factor 1.04 for a unit increment in the insulin level. In other word, the odds of occurrence of diabetes multiplied by 2.84 with one unit increment in diabetes pedigree function.

The coefficient value 0.83 of the bmiGrpobess indicates that the log odds of having diabetes increased by the 0.83 among the obsess women. In other word, the estimated odds of occurrence of diabetes multiplied by exp(0.83)=2.08 due the obesity factor in women. Similarly, we can infer regression coefficient for other variables.

Following this, we observed the effect of interaction term. The coefficient for Insulin:age_grpage$>= 50$ is 0.005 and p-value for Wald test statistics is slightly less than the 0.05 significance level. It indicates that the insulin level and having the age greater than 50 have combinatorial impact on the success probability. Hence, regression coefficient for Insulin will be (-0.003+ (-0.19*0.005)) = -0.00276 considering effect of ageGrp on it. Result shows that the effect of insulin level decreases by factor -0.19*0.004 = 0.0007 for a woman having age greater than 50. Thus, We can say that the log odds of occurrence of diabetes decrease by a factor 0.0027 for a unit increase in insulin level considering woman age is greater than 50.

Table 2: Odd ratio with 95% confidence Interval			
Variable	OR	Lower CI	Upper CI.
(Intercept)	0.0014	0.0001	0.0088
Pregnancies	1.084	1.008	1.167
Glucose	1.039	1.030	1.048
Insulin	0.99	0.9942	0.99
DiabetesPedigreeFunction	2.84	1.524	5.400
bmiGrpnormal	0.206	0.034	1.72
bmiGrpoverwt	0.73	0.156	5.350
bmiGrpobess	2.30	0.51	16.44
ageGrpage50	1.21	0.640	2.27
ageGrpage> 50	0.80	0.34	1.81
Insulin:ageGrpage50	1.00	0.99	1.007
Insulin:ageGrpage> 50	1.005	1.00	1.0114

As we can see from the table, odd ratio corresponding to pregnancies is 1.084 with 95% interval [1.008, 1.167]. This implies that if we fixed the other variables, increase in one unit of pregnancy will increase the odds of occurrence of diabetes by 0.08. Odd ratio corresponding to bmiGrpobess is 2.30 with 95% confidence interval [0.51 16.44]. This shows that the odd of having diabetes increase by factor 2.30 for obess woman. However, the confidence intervals values shows less precision in the estimates of the odd ratio.

Next, we used the model to predict the probabilities for new instances. Following graph shows the predicted probabilities along with their confidence intervals. The confidence intervals for the estimated probabilities are wide. This indicates the lack in precision while estimating the probabilities.

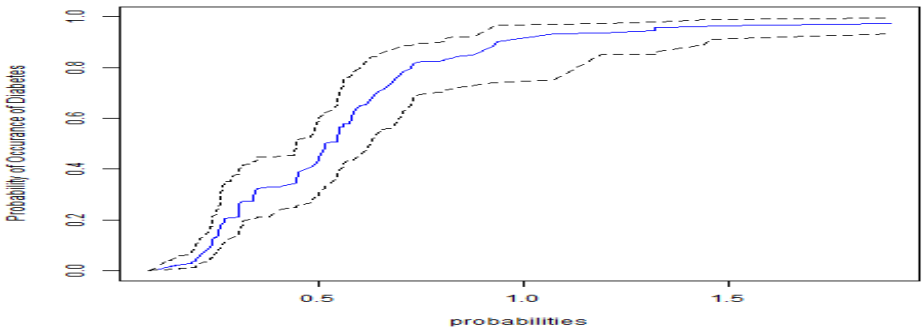


Figure 2: Predicted Probabilities with their 95% confidence interval

Conclusion and Discussion

To summarize, our response, occurrence of diabetes was binary. Therefore, we assumed binomial distribution to response and used logit function to model it. We used backward and subset selection approach to find the model with optimal set of predictors. In order to compare their performances, we used global fitness test - Pearson test statistics, compared their residual deviance. We used variable selection methods two find the optimal model. However, this approach doesn't always guarantee that the model, we selected is the best model. We used model 2 with interaction term as our final selected model and interpreted their their results. We found the increased diabetes pedigree function and glucose level are important predictors to identify the occurrence of diabetes. We used fitted model to predict the unknown frequency and found around 75% prediction accuracy. We can further use ROC estimate to measure the predictive power of fitted model.

Technical Description

I. Logistic Regression Model

Logistic Regression Model is generalized linear model which is used to model the response with binomial distribution.
For this glm model, we model the success probability of success i.e. p(y=1).
Thus, GLM components are as follows:
1. Random Component $= E(\pi(x)) = Y$
2. Systematic component $= \alpha + \sum_{i=1}^{p} \beta_i X_i$
3. Link function : $logit(\pi) = log(\frac{\pi}{1-\pi})$

$$logit([\pi(x)]) = log(\frac{\pi}{1-\pi}) = \alpha + \beta_1 X.....(1.2)$$

We take logit as link function which guarantees that values will lie between 0 and 1.

In order to find the probability, we take exponential on both side. After some calculation, we estimate the probability of success as follows:

$$\pi(y) = \frac{e^{\alpha + \sum_{i=1}^{p} \beta_i x_i}}{1 - e^{\alpha + \sum_{i=1}^{p} \beta_i X_i}}$$

Using this, we can assume that $Y = 1|x$ is binomial random variable with success probability $\pi(x)$

Next we use Maximum likelihood approach to find the unknown parameter β
p.d.f. for the Binomial distribution is given as follows:

$$P(Y = y_i) = \binom{n}{y}\pi^y(1-\pi)^{1-y}.....y = 0, 1, ...n$$

where $\pi(y) = \frac{e^{\alpha + \sum_{i=1}^{p} \beta_i x_i}}{1 - e^{\alpha + \sum_{i=1}^{p} \beta_i X_i}}$

$$L(\pi) = \prod_{i=1}^{n} \pi^y(1-\pi)^{1-y}$$

$$L(\pi) = \prod_{i=1}^{n} \pi^y . \prod_{i=1}^{n} (1-\pi)^{1-y}$$

$$L(\pi) = \prod_{i=1}^{n} [\frac{exp(\alpha + \beta X)}{1 + exp(\alpha + \beta x)}]^{Y_i} [\frac{1}{1 + exp(\alpha +)}]^{1-y_i}$$

We then take log on both side of the equation and find the derivative of likelihood function to find MLE. We can solve derivative of the likelihood function to get the value of unknown parametersβ

II. Interpretation of Logistic Regression Model

Consider a very simple Logistic regression model as follows:

$$logit(\pi^{(x)}) = \alpha + \beta_1 X - - - 2$$

$$\frac{\pi(x)}{1-\pi x} = e^{\alpha+\beta x}$$
$$\frac{\pi(x)}{1-\pi x} = e^{\alpha}e^{\beta x}$$

It means Odd increases multiplicativly with x.
Every unit increase in x leads to an increase in the odd of e^{β}
One unit increment of x is equals to:

$$\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}} = e^{\beta}$$

When $\beta = 0, e^{\beta} = 1$. Therefore odd do not change with x.

III. Residual Deviance

Residual Deviance of model calculated as below:
Residual deviance = (deviance of estimated model) - (deviance of ideal model where predicted values equals to the observed model)

IV. Global Fit Statistics

We use Global-Fit Statistics to assess how well model fits the observed data.
Null hypothesis for these statistic is $H_0 =$ The model fits the data and lack of fit is not statistically large
Pearson "goodness-of-fit" is:

$$\chi^2 = \sum_{i=1}^{N} \frac{y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

where,
$y_i =$ observed count
$\mu_i =$expected count
χ^2 is approximately Chi-square distributed with degrees of freedom(residual df)=number of counts - number of model parameter

IV. Pearson Residual

Pearson residual= $\frac{r_i}{\sqrt{\hat{\mu_i}}}$
To control the over-dispersion in the model, we use dispersion parameter ϕ and calculate Pearson residual as below: $p_i^c = \frac{r_i}{\sqrt{\phi\hat{\mu}_i}}$
Where,
$\phi = \frac{1}{n-k}\sum_{i=1}^{n}\frac{r_i^2}{\mu_i}$, $r_i =$ residual = $y_i - \hat{\mu i}$, $\mu_i = e^{\hat{\beta_0}+\beta_1\hat{X}_1 i+..+\beta_k\hat{X}_{ki}}$

V. Wald Statistic to test Significance of coefficients

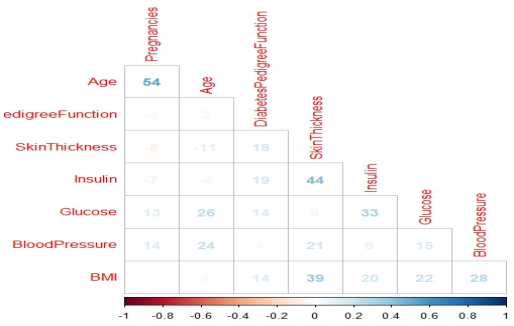
For Wald Statistics, we consider $H_0 : \beta = 0$ Vs $H_a = \beta \neq 0$

$$z = \frac{(\hat{\beta} - \beta_0)}{ASE} = \frac{(\hat{\beta})}{ASE}$$

Wald Statistic $= z^2 \approx \chi^2$
Where, $\hat{\beta} =$ estimate value of β , $\beta_0 =$ observed value of β, $ASE =$ Asymptotic Standard Error
We can use both z and z^2 for the Hypothesis testing.
We reject null hypothesis when $p(z) < 0.05$ and conclude variable is significant at 0.05 significance level.

5 Appendix

I. Correlation Plot



II. Summary Statistics

```
summary(diabetes_dat[,c(1,2,3,4,5,6,8,7)])
```

```
 Pregnancies        Glucose      BloodPressure    SkinThickness
Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
   Insulin           BMI             Age        DiabetesPedigreeFunction
Min.   :  0.0   Min.   : 0.00   Min.   :21.00   Min.   :0.0780
1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:24.00   1st Qu.:0.2437
Median : 30.5   Median :32.00   Median :29.00   Median :0.3725
Mean   : 79.8   Mean   :31.99   Mean   :33.24   Mean   :0.4719
3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:41.00   3rd Qu.:0.6262
Max.   :846.0   Max.   :67.10   Max.   :81.00   Max.   :2.4200
```

III. Data Summary

```
summary(diabetes_dat[,c(1,2,3,4,5,6,8,7)])
```

```
 Pregnancies        Glucose      BloodPressure    SkinThickness
Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
   Insulin           BMI             Age        DiabetesPedigreeFunction
Min.   :  0.0   Min.   : 0.00   Min.   :21.00   Min.   :0.0780
1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:24.00   1st Qu.:0.2437
Median : 30.5   Median :32.00   Median :29.00   Median :0.3725
Mean   : 79.8   Mean   :31.99   Mean   :33.24   Mean   :0.4719
3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:41.00   3rd Qu.:0.6262
Max.   :846.0   Max.   :67.10   Max.   :81.00   Max.   :2.4200
```

III. Hosmer-Lemenshow's Lack-of-fit Test Expected and Fitted values for 10 groups

	y0	y1	yhat0	yhat1
[0.0015,0.0404]	64.000000	2.000000	64.572368	1.427632
(0.0404,0.0827]	62.000000	3.000000	60.868407	4.131593
(0.0827,0.133]	61.000000	4.000000	57.867769	7.132231
(0.133,0.186]	51.00000	14.00000	54.54745	10.45255
(0.186,0.255]	52.00000	13.00000	50.53421	14.46579
(0.255,0.341]	44.00000	21.00000	46.07578	18.92422
(0.341,0.436]	42.00000	23.00000	40.04645	24.95355
(0.436,0.61]	31.00000	34.00000	31.94343	33.05657
(0.61,0.811]	16.00000	49.00000	19.24206	45.75794
(0.811,0.978]	10.000000	56.000000	7.302065	58.697935

6 References

1. Dr. Cheng Peng. (Fall 2016 - STA 588) Notes on Generalized Linear Models and Logistic Regression
2. James, G. et al. (2013). An introduction to statistical learning (Vol. 112). New York: Springer
3. https://onlinecourses.science.psu.edu/stat501/node/374 Estimators and Tests'
4. https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes