PM566 Final Project

Jun Lee

Table of Contents

Introduction

Diabetes is a complex metabolic syndrome and its involvement in various diseases is manifold with varying manifestations and different clinical symptoms and prognosis. Although being probably the most important risk factor, diabetes is often considered an "accompanying comorbidity" for cardiac or peripheral artery disease, hypertension, or stroke. An interesting study demonstrated that the risk of sudden cardiac death is at least 2-times higher in patients who are diabetic compared with those who are nondiabetic, regardless of the extent of cardiac dysfunction or symptoms of heart failure. In contrast, the risk of nonsudden cardiac death was not significantly different between the 2 groups.

Analyzing diabetes is beneficial as it could help prevent tragedic illnesses such as sudden cardiac death, the leading cause of death in the United States. The object of this study is to find out if diabetes risk prediction is possible based on family history and genetic factors along with other risk factors.

This study is focused on 1) whether Diabetes Pedigree Function is truly associated with diabetes and if so, what risk factors have a significant relationship with it. 2) Creating effective diabetes prediction model. The association between Diabetes Pedigree Function and disease status would reveal to what extent does the genetic factors affect diabetes. The dataset did not separate type 1 diabetes and type 2 diabetes. Regardless of the type of diabetes, however, the dataset has sufficient predictors to evaluate the hypothesis because both types of diabetes are caused by impaired glucose metabolism. Diabetes Pedigree Function, which is the key variable, can be assessed as family history is one of the important risk factors for both types of diabetes. In the analysis, important indicators of diabetes will be evaluated together.

Data description

- Lipotoxicity (1-17): lipotoxicity is a metabolic syndrome that results from the accumulation of lipid intermediates in non-adipose tissue, leading to cellular dysfunction and death.

- Glucose (mmol/L): blood glucose level obtained by measuring plasma glucose concentration at 2 hours in an oral glucose tolerance test.

- BloodPressure (mm Hg): the pressure of the blood in the circulatory system.

- SkinThickness (mm): skin thickness is primarily determined by collagen content and is increased in insulin-dependent diabetes mellitus.

- Insulin(μU/ml): insulin is an anabolic hormone that promotes glucose uptake.

- BMI: body mass index (BMI) is a person's weight in kilograms divided by the square of height in meters.

- DiabetesPedigreeFunction (0:1 value generated from familial diabetes history/risk): diabetes pedigree function provides "a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject." It generally provides scores of the likelihood of diabetes based on family history. The DPF uses information from parents, grandparents, siblings, aunts and uncles, and first cousins. It provides a measure of the expected genetic influence of affected and unaffected relatives on the subject's eventual diabetes risk.

- Age: age of the individual.

- Outcome: diabetes test result (0 = Non-diabetic, 1 = Diabetic).

Key questions

Is there a significant difference in values of diabetes risk factors for those who have diagnosed with diabetes and those who are not? Is Diabetes Pedigree Function significantly associated with the onset of diabetes and other risk factors? Could these risk factors provide a reliable prediction of individual's diabetes?

Methods

The dataset used in this study was acquired from Harvard Dataverse. Among several predictors, Diabetes Pedigree Function was a particularly interesting attribute in the dataset. It provided some data on diabetes history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence became the starting point of an idea that the hereditary risk could be used for the risk stratification or prediction of onset of diabetes.

Reference: P. Rodgers, Judith, 2020, "Diabetes Mellitus "Comorbidity" or an Important Risk Factor?", https://doi.org/10.7910/DVN/JAW6AX, Harvard Dataverse, V2, UNF:6:QOwrJ53n2F5fMC+wb4ADVA== [fileUNF]

Analysis through assessing various plots, tables and graphs was performed to identify association between Diabetes Pedigree Function and diabetes test outcome, including the examination of the effect of lipotoxicity, glucose level, blood pressure, skin thickness, insulin level, BMI, and age. The data was cleaned by replacing extreme values to 'NA's, shortening variable names, and creating new factor variable for better analysis. Skim function from skimr package was used to explore data. Dim, head and tail, summary, and

table functions were used to check detailed observations. Age was stratified into four age groups(20-29, 30-39, 40-49, and 50+) for better understanding of relationship with diabetes. Outcome was binomial variable(0, 1) and it was transformed into factor variable (Non-diabetic and Diabetic). For prediction model, logistic regression was performed with grouped smooth method and LOESS smoothing method conducted for linearity assumption evaluation. Influential outliers and model fit was checked and the statistics including accuracy, sensitivity, and specificity were displayed with graphs. ROC curve and area under the curve was measured for the discrimination ability of the model.

Exploratory Data Analysis

- The total dataset includes 768 observations with 9 vaiables.

- Lipotoxicity is right skewed and ranged from 0 to 17 with mean of 3.85.

- Glucose level is not skewed to either side and ranged from 0 to 199 with mean of 121.

- Blood Pressure is also not skewed and ranged from 0 to 122 with mean of 69.1.

- Skin Thickness right skewed and ranged from 0 to 99.

- Insulin level is extremly right skewed and ranged from 0 to 846.

- BMI is fairly normally distributed and ranged from 0 to 67.1.

- Diabetes Pedigree Function is right skewed and ranged from 0.078 to 2.42.

- Age is right skewed and ranged from 21 to 81. Right skewed age data displays more than 50% of observations are in the age group of 20 to 29. This data is more focused on young population.

- Outcome has binary results with mean of 0.349 which means 34.9% of all observations have diabetes positive outcome.

Data summary

Name	diabetes
Number of rows	768
Number of columns	9

_____

Column type frequency:

numeric	9

_____

Group variables	None

Variable type: numeric

| skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Lipotoxicity	0	1	3.85	3.37	0.00	1.00	3.00	6.00	17.00	▆▁▃▁
Glucose	0	1	120.89	31.97	0.00	99.00	117.00	140.25	199.00	▁▁▇▆
BloodPressure	0	1	69.11	19.36	0.00	62.00	72.00	80.00	122.00	▁▁▇▆▁
SkinThickness	0	1	20.54	15.95	0.00	0.00	23.00	32.00	99.00	▇▆▃▁▁
Insulin	0	1	79.80	115.24	0.00	0.00	30.50	127.25	846.00	▇▁▁▁▁
BMI	0	1	31.99	7.88	0.00	27.30	32.00	36.60	67.10	▁▃▇▃▁
DiabetesPedigreeFunction	0	1	0.47	0.33	0.08	0.24	0.37	0.63	2.42	▆▅▁▁▁
Age	0	1	33.24	11.76	21.00	24.00	29.00	41.00	81.00	▆▅▁▁▁
Outcome	0	1	0.35	0.48	0.00	0.00	0.00	1.00	1.00	▇▁▁▁▆

Data Cleaning

- 0 values in all risk factors are speculated as an absence of the specific test outcome because there were significant difference in values between all 0s and the next minimum values of each variable. All of them are edited to "NA"s.

- Variable names are renamed into lowercase letters with shorter length.

- Created age groups (20-29, 30-39, 40-49, 50+) to compare proportion of the diabetes by age groups.

Results

Histogram of Diabetes outcome by age groups

- Age is a significant risk factor for diabetes. As age is a well-known confounder of most diseases, it could also play a role as a confounder when generating a prediction model. Through this histogram, I can confirm that age affects the onset of diabetes.

Diabetes Outcome by Age Groups
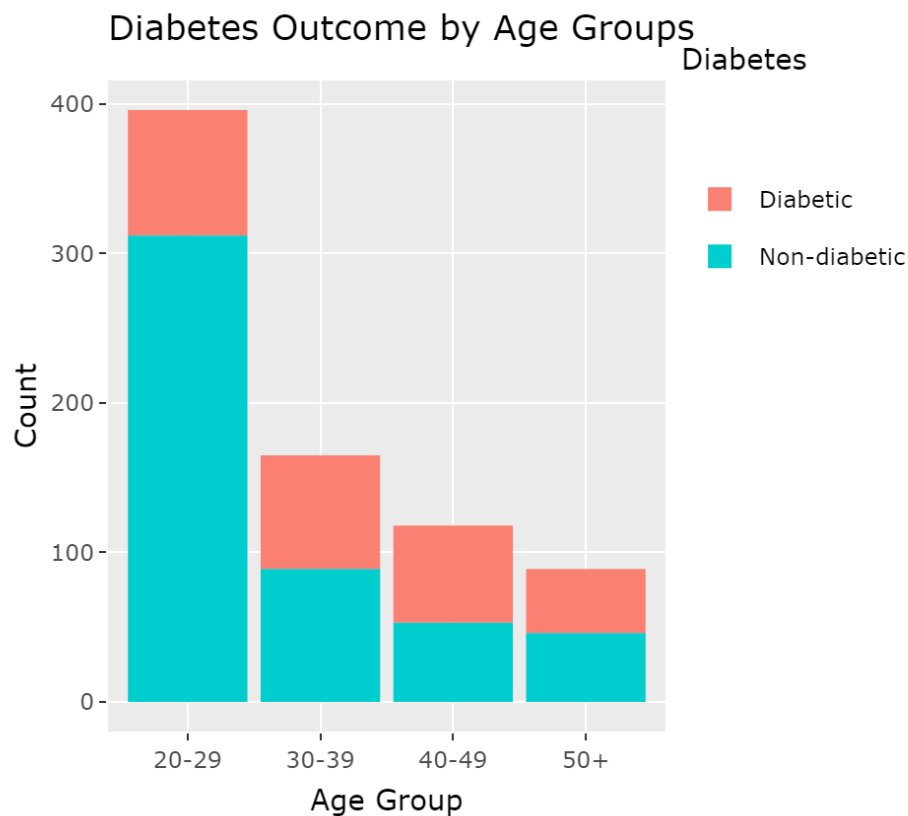
Mean Values of Each Risk Factors by Outcome Group

- All of the predictors are showing some differences in mean values by diabetes outcome, meaning that these predictors can be utilized for a prediction model.

- Diabetes test outcome: 0 (Non-diabetic), 1 (Diabetic).

Mean of Each Risk Factors by Outcome Group

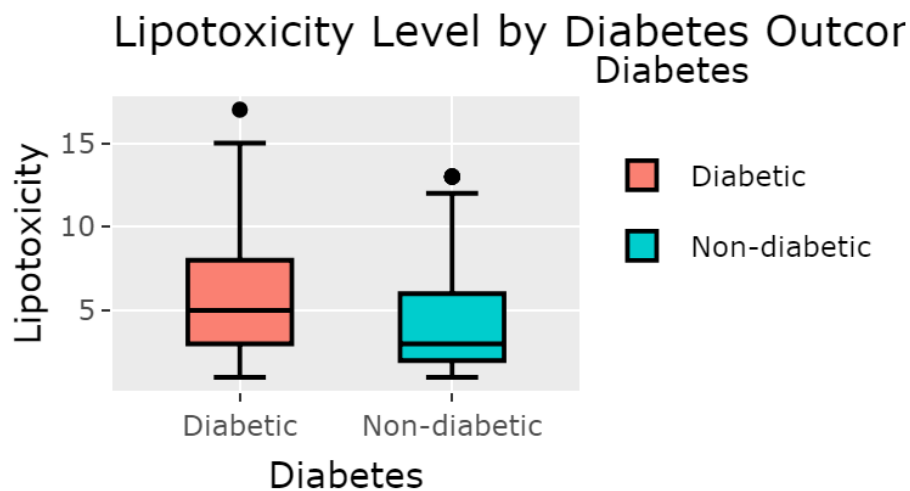| outcome	DiabetesPedigreeFunction	Lipotoxicity	Glucose	BloodPressur	SkinThickness	Insulin	BMI	Age
0	0.43	3.86	110.64	70.88	27.24	130.29	30.86	31.19
1	0.55	5.67	142.32	75.32	33.00	206.85	35.41	37.07

Box Plot of each risk factors by outcome

- Among the predictors, Glucose Level, Insulin level, and BMI showed significant differences in mean values by outcome group.
- Insulin level is a direct indicator of discernment between type 1 and 2 diabetes. This predictor can be treated differently by the type of diabetes. In this dataset, it is

presumed that there were more type 2 diabetes in the patient group as the result shows that mean insulin level is higher in the diabetic group.

- Although test for measuring lipotoxicity is not common, lipotoxicity is showing a meaningful gap between the diabetes and non-diabetes group.
- Skin thickning is a symptom detected from patients with insulin-dependent diabetes mellitus (IDDM). It means this data only applies to type 1 diabetes.
- Blood Pressure does not show significant difference by diabetes outcome.

Lipotoxicity

Glucose Level

Blood Pressure

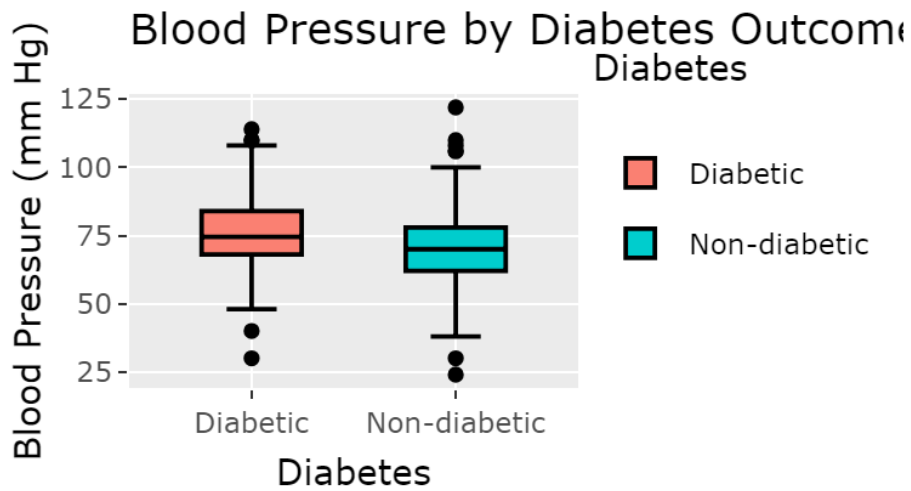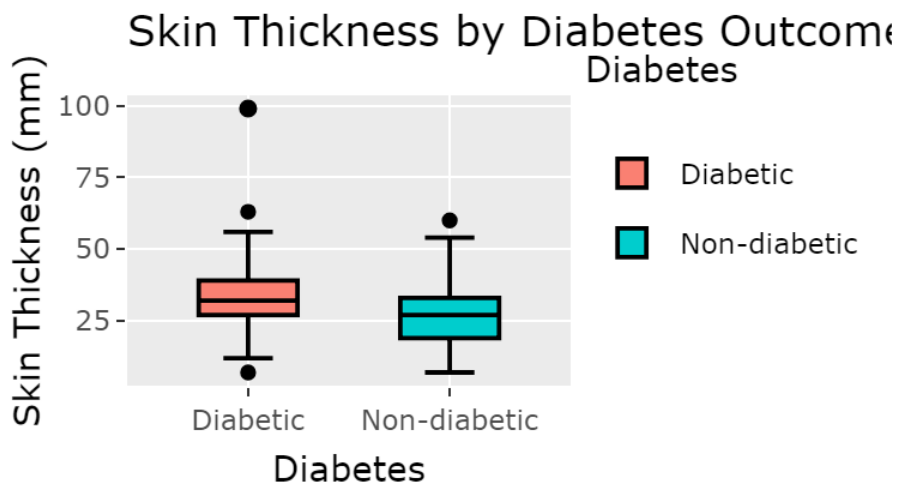Blood Pressure by Diabetes Outcome

Skin Thickness
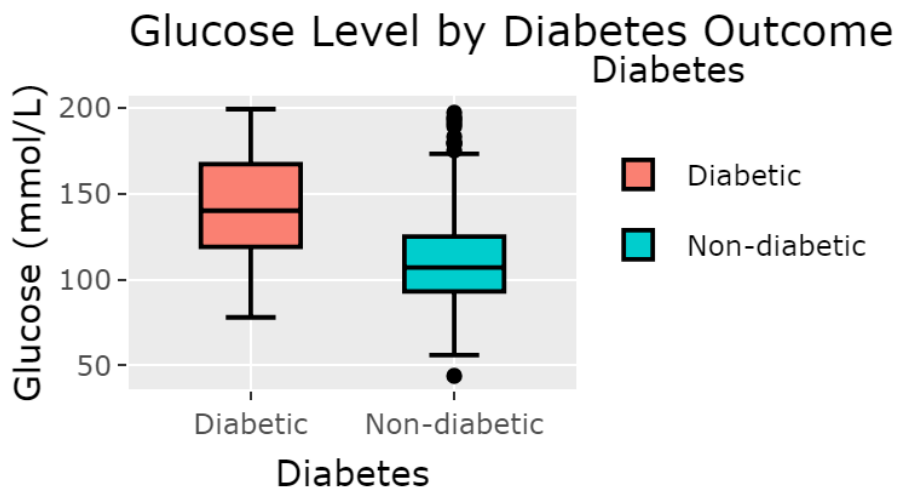
Skin Thickness by Diabetes Outcome

Insulin Level
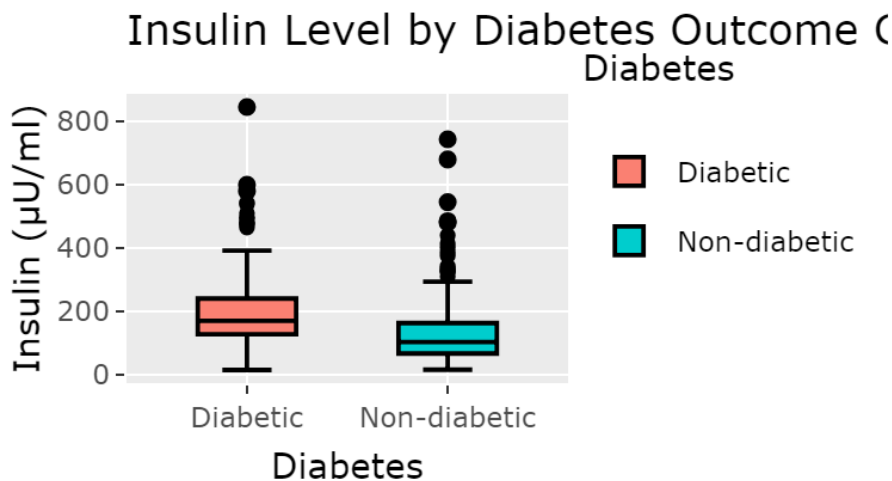
BMI
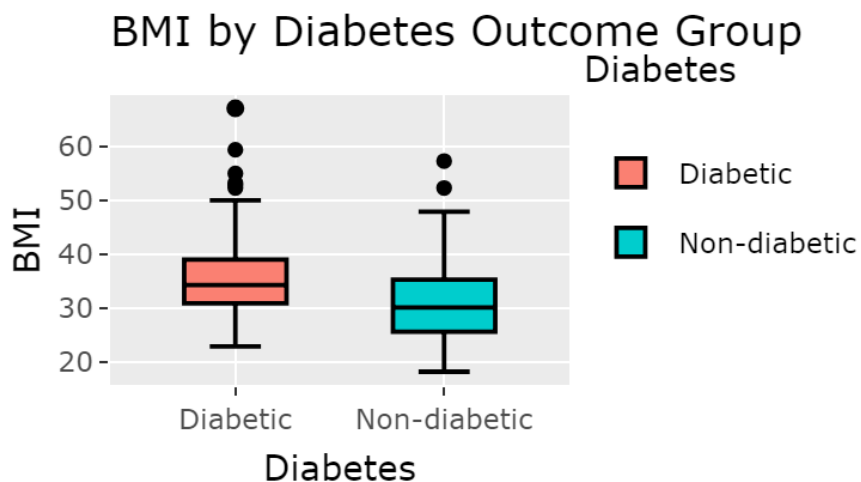
Plots of Diabetes Outcome by Diabetes Pedigree Function

- The histogram shows that Diabetes Pedigree Function follows Poisson distribution. Poisson regression could be used for further analysis.
- The proportion of diabetic outcome increases over Diabetes Pedigree Fucgion. From dpf 0.24 to 0.48 section, proportion of diabetic vs non-diabetic is approximately 1:2. In 0.56 to 0.72 section, the ratio is 2:3 and in 0.8 and over section, the ratio become close to 1:1.
- Boxplot also shows that there is a significant difference in Diabetes Pedigree Function value between diabetic and non-diabetic group. There is more research needed for the outliers without diabetes.

Diabetes Pedigree Fuction Distributio
Diabetes

Diabetic / Non-diabetic

Diabetes Pedigree Fuction by Diabet
Diabetes
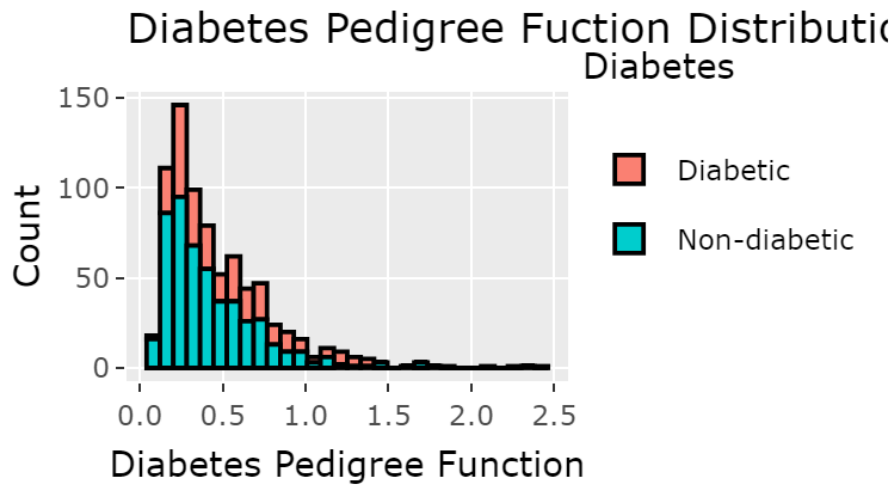
Diabetic / Non-diabetic
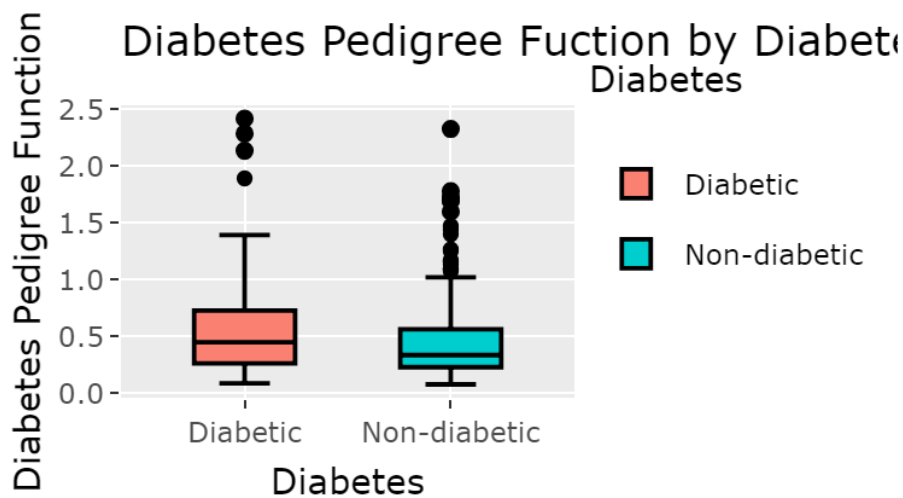
Scatter Plot Graph of Significant Risk Factors vs Diabetes Pedigree Function by Diabetes Outcome Group

- Glucose level and diabetes pedigree fuction interact well to distinguish diabetic section and non-diabetic section.
- Insulin level is not effective on marking off diabetic section in this scatter plot. However, it could be different when it only applies to the dataset of patients who entirely has one of two types of diabetes.
- BMI is also a good partner of diabetes pedigree function.

- Upper right side of the lipotoxicity vs DPF graph displays more diabetic outcome as both lipotoxicity and diabetes pedigree function are high.

Glucose vs DPF

Glucose Level vs Diabetes Pedigree Function by oucome

Insulin vs DPF

Insulin Level vs Diabetes Pedigree Function by oucome

Diabetes Outcome:Non-diabetic
Diabetes Pedigree Function: 1.268
Insulin Level: 59

Diabetes Outcome:Diabetic
Diabetes Pedigree Function: 1.321
Insulin Level: 58

BMI vs DPF

BMI vs Diabetes Pedigree Function by oucome

Lipotoxicity vs DPF
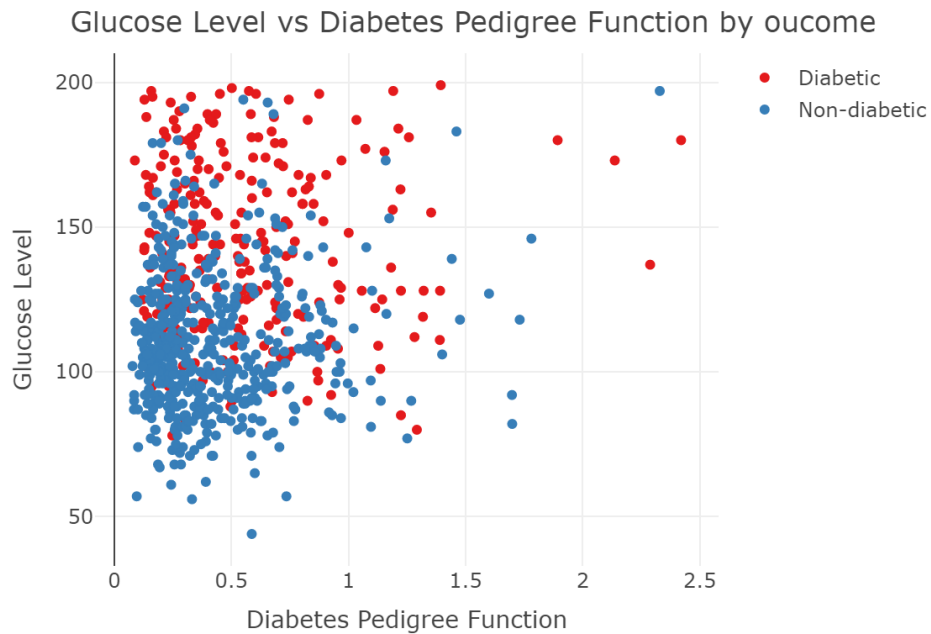
Lipotoxicity vs Diabetes Pedigree Function by oucome
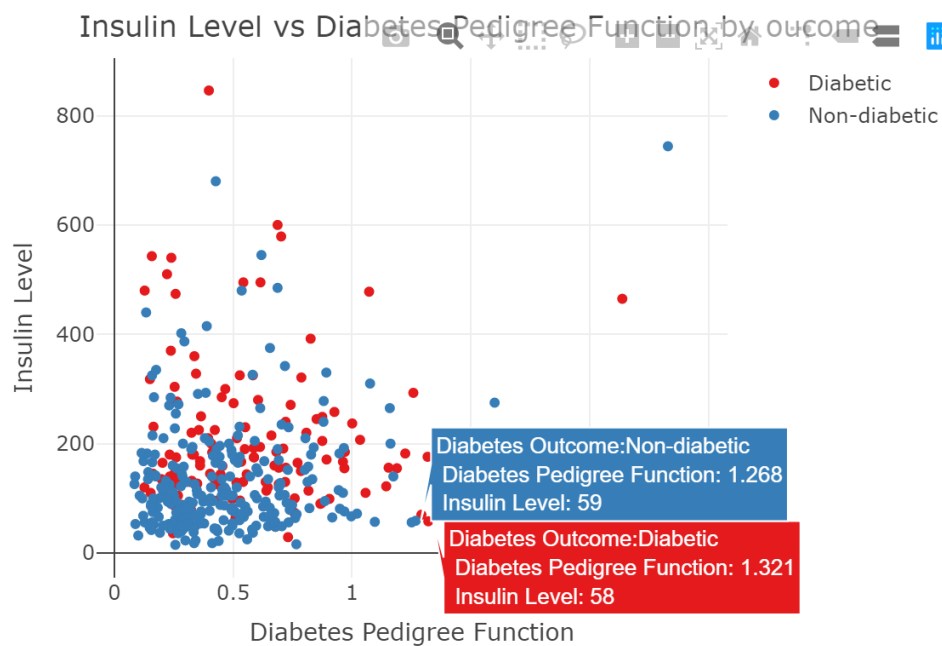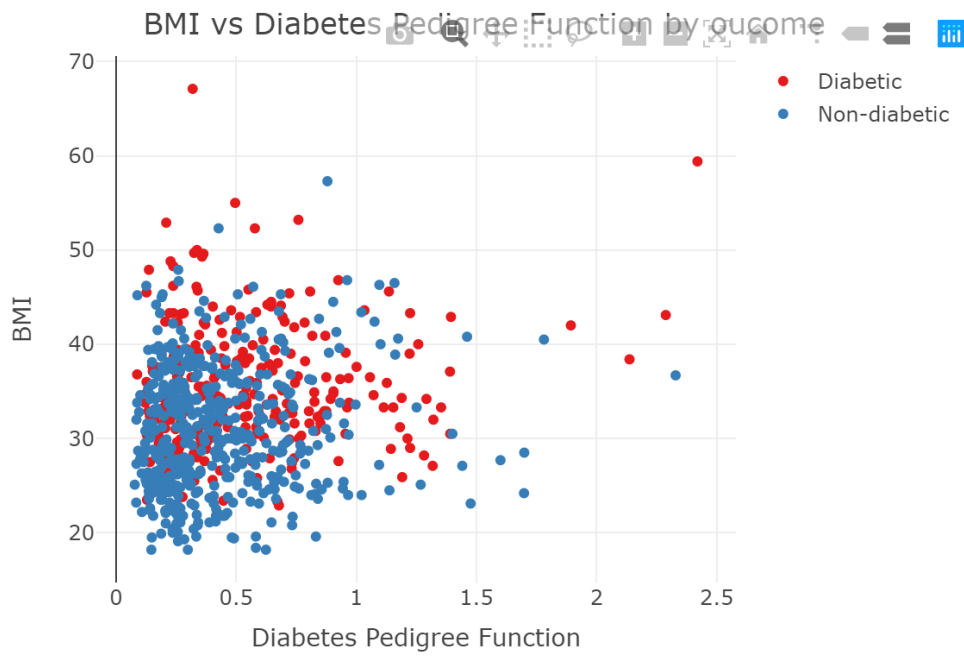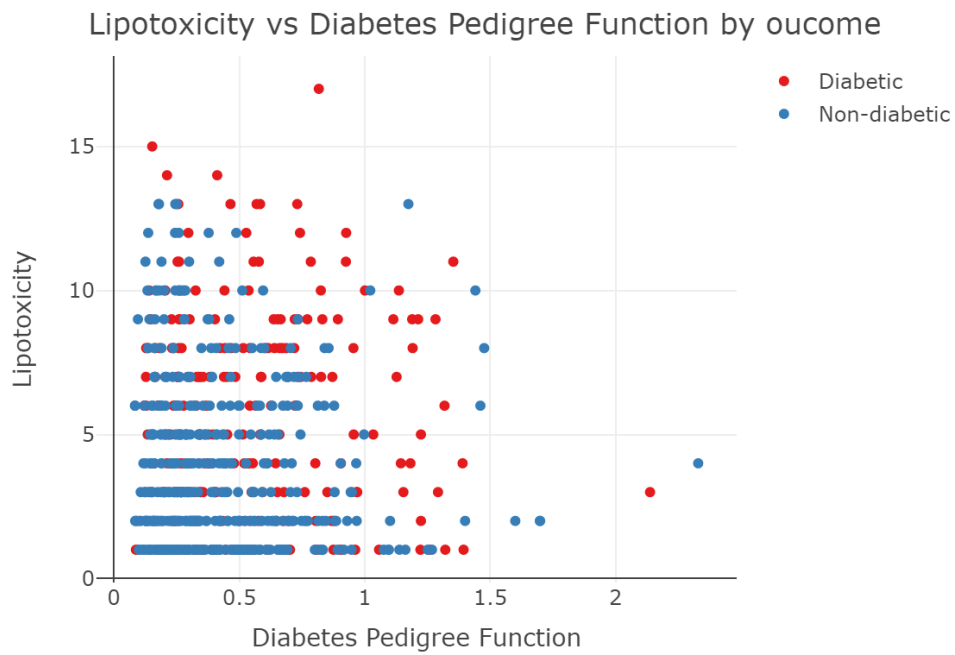
Conclusion

Overall, most of the predictors displayed differences in mean values when it compares between diabetic and non-diabetic groups. The interesting predictor was diabetes pedigree function because, unlike other risk factors, DPF could be measured by relatives history and genetic data. This is the only predictor that can be obained by external sources other than individual's biological test result. In the analysis, diabetes pedigree function showed its association with the onset of diabetes. Although it is hard to predict the risk of diabetes with diabetes pedigree fuction alone, the analysis showed a possibility of utilization of other risk factors combined with diabetes pedigree funtion in prediction model. In this study, glucose level and BMI provided evidence of significant association with onset of diabetes in conjunction of diabetes pedigree function. Insulin level and lipotoxicity also showed some association to be part of prediction model. Based on the visualized plot analysis, I performed logistic regression to build a prediction model. The dataset was separated into 70% training dataset and 30% test dataset. For the univariable analysis, multivariable fractional polynomial function was used to assess best fit model, and grouped smooth and LOESS smoothing method were used to check linearity assumption. Through this process, association between diabetes outcome and DPF, lipotoxicity, glucose level, blood pressure, skin thickness, insulin level, BMI, and age was evaluated. Overall, DPF, lipotoxicity, glucose level, and age did not violated linearity under logit setting. For the next step, glmulti function was used to obtain best multivariable model. All variables were included in the evaluation and the best model resulted in excluding blood pressure and skin thickness. It is not a surprising decision as I obtained similar result from previous visualized plot analysis. Through level 2 glmulti assessment, I was able to reveal interaction between insulin level and DPF, insulin level and lipotoxicity, and DPF and BMI. However, sim_slopes function indicated that interaction between insulin level and lipotoxicity, and DPF and BMI are falsely identified. The interaction between insulin level and DPF was scientifically plausible as DPF is direct genetic value from diabetes history and it is closely associated with insulin level at certain level. Age was assessed if it is a confounder. More than 10% of parameter estimate change in lipotoxicity after applying age in the model supported that it is a confounder. As a result, the final model was built as following.

$$\text{logit}(\pi) = \beta_0 + \beta_{dpf} x_{dpf} + \beta_{lip} x_{lip} + \beta_{glu} x_{glu} + \beta_{ins} x_{ins} + \beta_{bmi} x_{bmi} + \beta_{age} x_{age} + \beta_{ins} x_{ins} \beta_{dpf} x_{dpf}$$

Hosmer and Lemeshow goodness of fit test showed that the model has good fit by not objecting null hypothesis ($p > 0.05$). No influential outliers were identified in Pearson's residual plot. For prediction model, logistic regression was performed with grouped smooth method and LOESS smoothing method conducted for linearity assumption evaluation. Influential outliers and model fit was checked and the statistics including accuracy, sensitivity, and specificity were displayed with graphs. ROC curve and area under the curve was measured for the discrimination ability of the model. The statistics of the training dataset provided around 80% accuracy with around 60% sensitivity and 80% specificity. ROC curve was created and the area under the curve value was 0.87 which indicates good discriminative ability. Same methods were applied to the test data set. Similarly, around 80% accuracy and 0.87 AUC were obtained.

Through this study, I found out that prediction model for diabetes could be built with relevant predictors such as diabetes pedigree function, glucose level, insulin level, lipotoxicity, BMI, and age. Further study will be needed with larger dataset for better accuracy. Also, detailed analysis for prediction model to distinguish between type 1 and type 2 diabetes can be conducted with deeper depth of variety of predictor with larger dataset. Overall, this study can be a good starting point of further research.

Appendix

Prediction Model

Dataset Information

- The dataset includes 768 observations (500 non-diabetic, 268 diabetic). Training dataset accounts 70% of randomly selected observations and test dataset accounts rest of 30% of observations.

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  768
## 
## 
##            |          0 |          1 |
##            |------------|------------|
##            |        500 |        268 |
##            |      0.651 |      0.349 |
##            |------------|------------|
## 
## 
## 
## 
```

Univariate Analysis

- Before I build a preliminary main effects model, I performed univariable analaysis. All the variables in the dataset was analysed as there was no unappropriate predictor based on the table and plots assessment. Multivariable fractional polynomial was used to obtain a best fit model of each variable.

```
## Call:
## mfp::mfp(formula = outcome ~ fp(dpf), data = db_train, family = binomial)
## 
## 
```

```
## Deviance table:
##           Resid. Dev
## Null model    719.0036
## Linear model  701.7728
## Final model   701.7728
##
## Fractional polynomials:
##     df.initial select alpha df.final power1 power2
## dpf          4      1  0.05        1      1      .
##
##
## Transformations of covariates:
##      formula
## dpf I(dpf^1)
##
## Rescaled coefficients:
## Intercept      dpf.1
##    -1.109      1.183
##
## Degrees of Freedom: 548 Total (i.e. Null);  547 Residual
## Null Deviance:       719
## Residual Deviance: 701.8     AIC: 705.8

## Analysis of Deviance Table
##
## Model 1: y ~ meanx
## Model 2: y ~ factor(meanx)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       547     704.69
## 2       545     699.37  2   5.3233  0.06983 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
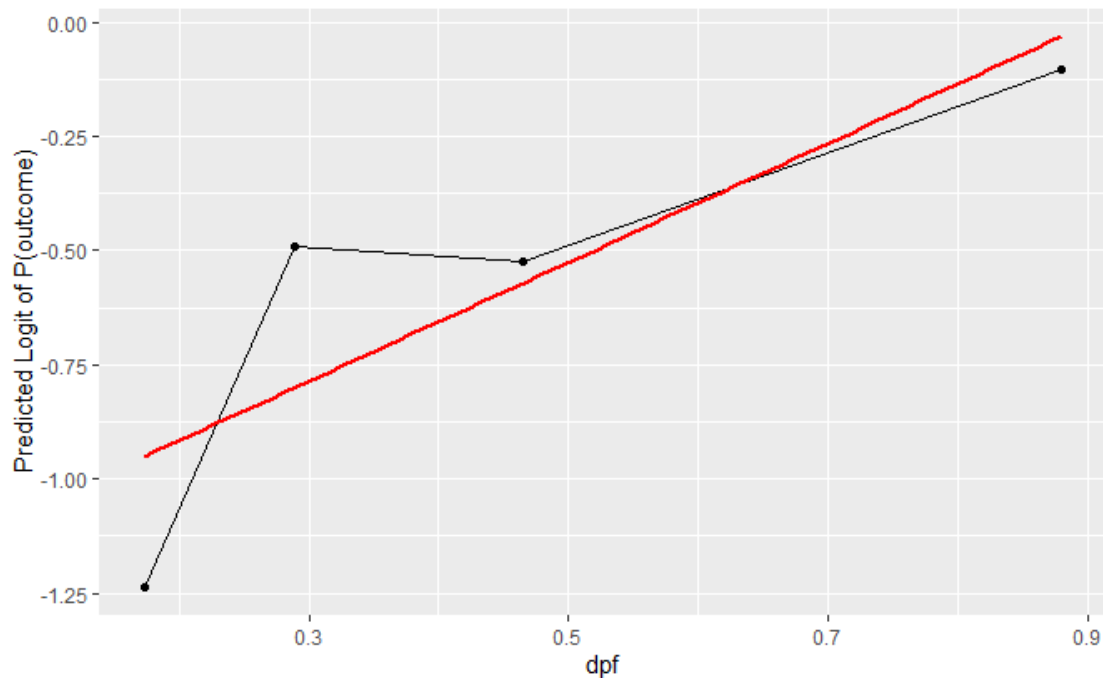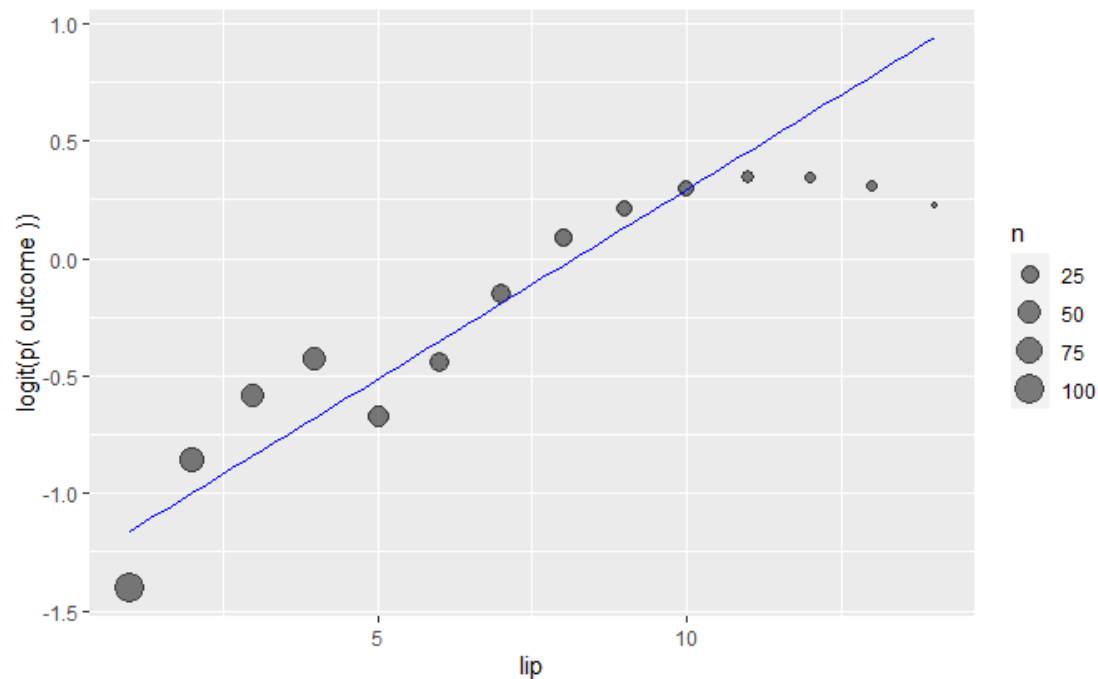
```
## 
## Call:
## glm(formula = outcome ~ dpf, family = binomial, data = db_train)
## 
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9092  -0.9222  -0.8368   1.3302   1.6231
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1094     0.1642  -6.758 1.40e-11 ***
## dpf           1.1832     0.2959   3.999 6.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 719.00  on 548  degrees of freedom
## Residual deviance: 701.77  on 547  degrees of freedom
## AIC: 705.77
## 
## Number of Fisher Scoring iterations: 4

##                   2.5 %      97.5 %
## (Intercept) -1.4376693 -0.7932477
## dpf          0.6160291  1.7788872

## Call:
## mfp::mfp(formula = outcome ~ fp(lip), data = db_train, family = binomial)
```

```
## 
## 
## Deviance table:
##            Resid. Dev
## Null model    616.9188
## Linear model  588.9274
## Final model   588.9274
## 
## Fractional polynomials:
##     df.initial select alpha df.final power1 power2
## lip          4      1  0.05        1      1      .
## 
## 
## Transformations of covariates:
##           formula
## lip I((lip/10)^1)
## 
## Rescaled coefficients:
## Intercept      lip.1
##   -1.3319     0.1634
## 
## Degrees of Freedom: 471 Total (i.e. Null);  470 Residual
## Null Deviance:        616.9
## Residual Deviance: 588.9      AIC: 592.9
```
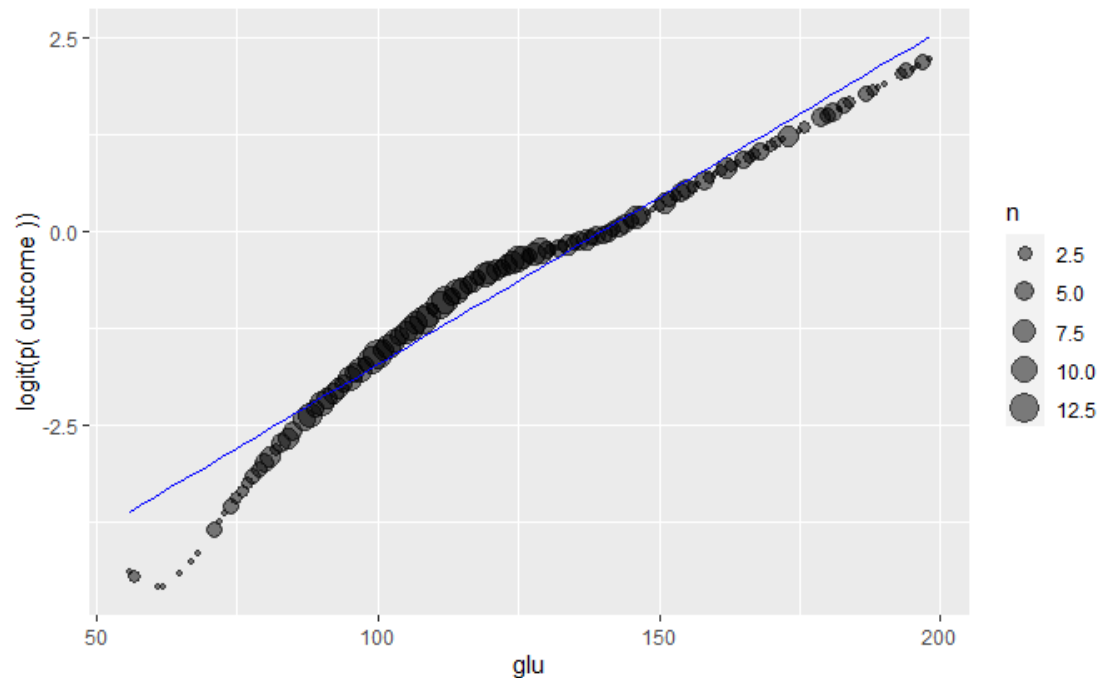


```
## 
## Call:
## glm(formula = outcome ~ lip, family = binomial, data = db_train)
## 
```
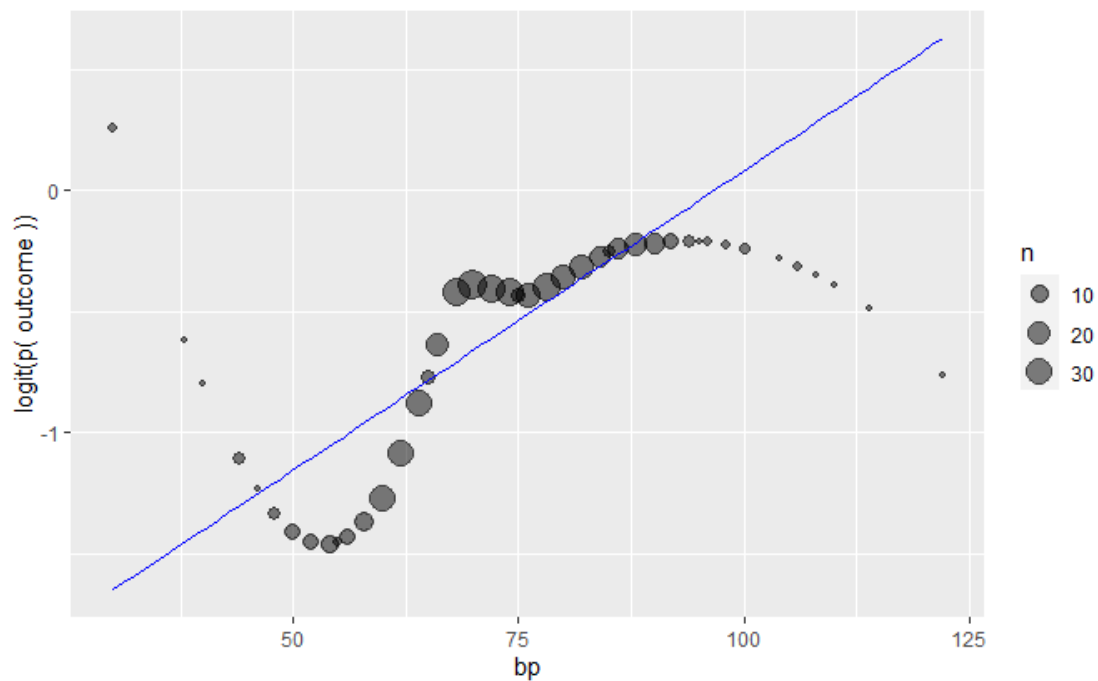
```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5271  -0.9061  -0.7358   1.2582   1.6966
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.33192    0.18038  -7.384 1.53e-13 ***
## lip          0.16343    0.03169   5.157 2.50e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 616.92  on 471  degrees of freedom
## Residual deviance: 588.93  on 470  degrees of freedom
##   (77 observations deleted due to missingness)
## AIC: 592.93
##
## Number of Fisher Scoring iterations: 4

##                  2.5 %      97.5 %
## (Intercept) -1.6928024 -0.9847578
## lip          0.1021083  0.2265587

## Call:
## mfp::mfp(formula = outcome ~ fp(glu), data = db_train, family = binomial)
##
##
## Deviance table:
##            Resid. Dev
## Null model   713.1413
## Linear model 563.1485
## Final model  563.1485
##
## Fractional polynomials:
##     df.initial select alpha df.final power1 power2
## glu          4      1  0.05        1      1      .
##
##
## Transformations of covariates:
##            formula
## glu I((glu/100)^1)
##
## Rescaled coefficients:
## Intercept      glu.1
##  -5.68215    0.04096
##
## Degrees of Freedom: 544 Total (i.e. Null);  543 Residual
## Null Deviance:       713.1
## Residual Deviance: 563.1     AIC: 567.1
```

```
## 
## Call:
## glm(formula = outcome ~ glu, family = binomial, data = db_train)
## 
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2246  -0.7598  -0.5048   0.8076   2.2660
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.682152   0.509127  -11.16   <2e-16 ***
## glu          0.040958   0.003945   10.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 713.14  on 544  degrees of freedom
## Residual deviance: 563.15  on 543  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 567.15
## 
## Number of Fisher Scoring iterations: 4
## 
##                   2.5 %      97.5 %
## (Intercept) -6.71664880 -4.71769925
## glu          0.03348255  0.04897181
```

```
## Call:
## mfp::mfp(formula = outcome ~ fp(bp), data = db_train, family = binomial)
##
##
## Deviance table:
##             Resid. Dev
## Null model     686.6991
## Linear model   674.9161
## Final model    674.9161
##
## Fractional polynomials:
##     df.initial select alpha df.final power1 power2
## bp           4      1  0.05        1      1      .
##
##
## Transformations of covariates:
##           formula
## bp I((bp/100)^1)
##
## Rescaled coefficients:
## Intercept        bp.1
##    -2.4499      0.0255
##
## Degrees of Freedom: 526 Total (i.e. Null);  525 Residual
## Null Deviance:        686.7
## Residual Deviance: 674.9      AIC: 678.9
```
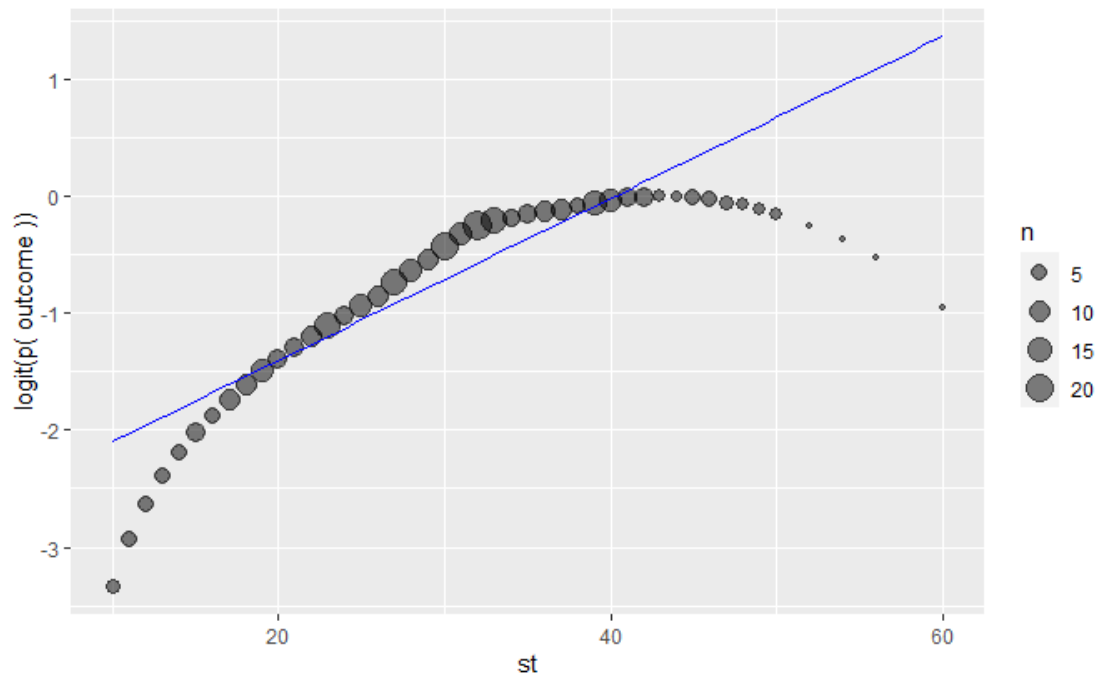


```
##
## Call:
```

```
## glm(formula = outcome ~ bp, family = binomial, data = db_train)
##
## Deviance Residuals:
##     Min      1Q   Median      3Q      Max
## -1.4680  -0.9496  -0.8191   1.3331   1.9261
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.449923   0.563501  -4.348 1.38e-05 ***
## bp           0.025502   0.007569   3.369 0.000753 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 686.70  on 526  degrees of freedom
## Residual deviance: 674.92  on 525  degrees of freedom
##   (22 observations deleted due to missingness)
## AIC: 678.92
##
## Number of Fisher Scoring iterations: 4

##                   2.5 %      97.5 %
## (Intercept) -3.57570625 -1.36267383
## bp           0.01084802  0.04057608

## Call:
## mfp::mfp(formula = outcome ~ fp(st), data = db_train, family = binomial)
##
##
## Deviance table:
##             Resid. Dev
## Null model    500.5365
## Linear model  472.3247
## Final model   472.3247
##
## Fractional polynomials:
##    df.initial select alpha df.final power1 power2
## st           4      1  0.05        1      1      .
##
##
## Transformations of covariates:
##         formula
## st I((st/10)^1)
##
## Rescaled coefficients:
## Intercept      st.1
##  -2.47476   0.06021
##
## Degrees of Freedom: 389 Total (i.e. Null);  388 Residual
```

```
## Null Deviance:        500.5
## Residual Deviance: 472.3      AIC: 476.3
```
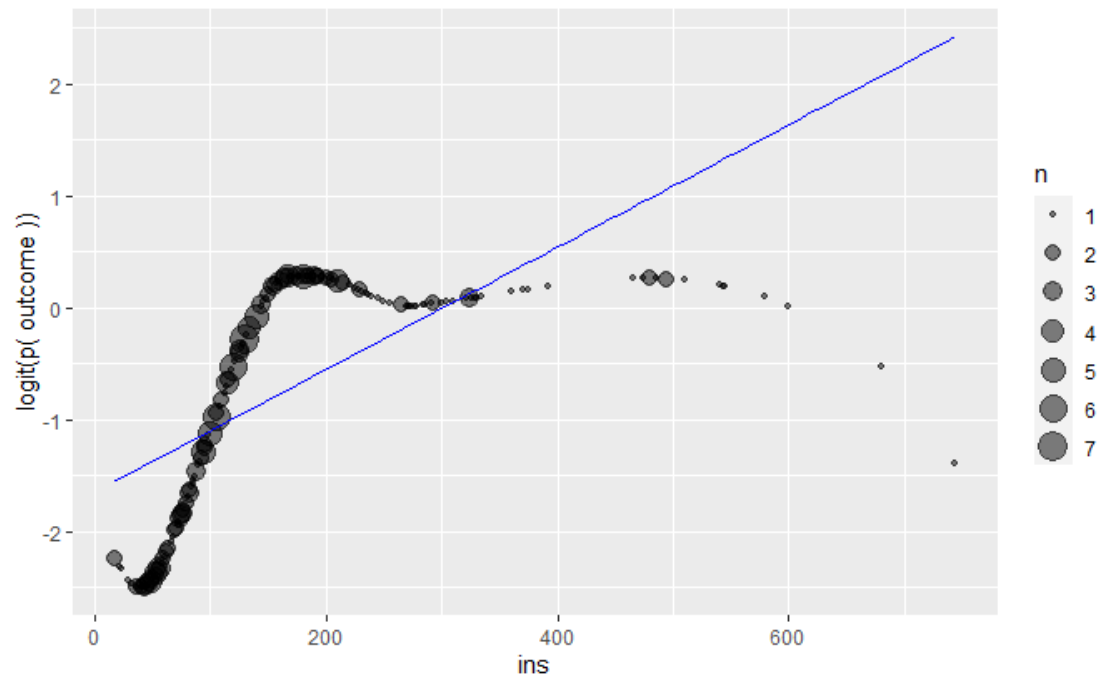


```
##
## Call:
## glm(formula = outcome ~ st, family = binomial, data = db_train)
##
## Deviance Residuals:
##     Min      1Q   Median      3Q      Max
## -1.6827  -0.9097  -0.6848   1.2316   1.9293
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.47476    0.38332  -6.456 1.07e-10 ***
## st           0.06021    0.01187   5.073 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 500.54  on 389  degrees of freedom
## Residual deviance: 472.32  on 388  degrees of freedom
##   (159 observations deleted due to missingness)
## AIC: 476.32
##
## Number of Fisher Scoring iterations: 4
```

```
##                     2.5 %      97.5 %
## (Intercept) -3.24976485 -1.74398014
## st           0.03742337  0.08405596

## Call:
## mfp::mfp(formula = outcome ~ fp(ins), data = db_train, family = binomial)
##
##
## Deviance table:
##             Resid. Dev
## Null model   365.8654
## Linear model  339.7208
## Final model   321.331
##
## Fractional polynomials:
##     df.initial select alpha df.final power1 power2
## ins          4      1  0.05        2   -0.5      .
##
##
## Transformations of covariates:
##                formula
## ins I((ins/100)^-0.5)
##
## Rescaled coefficients:
## Intercept       ins.1
##      2.18      -31.02
##
## Degrees of Freedom: 280 Total (i.e. Null);  279 Residual
## Null Deviance:        365.9
## Residual Deviance: 321.3     AIC: 325.3
```
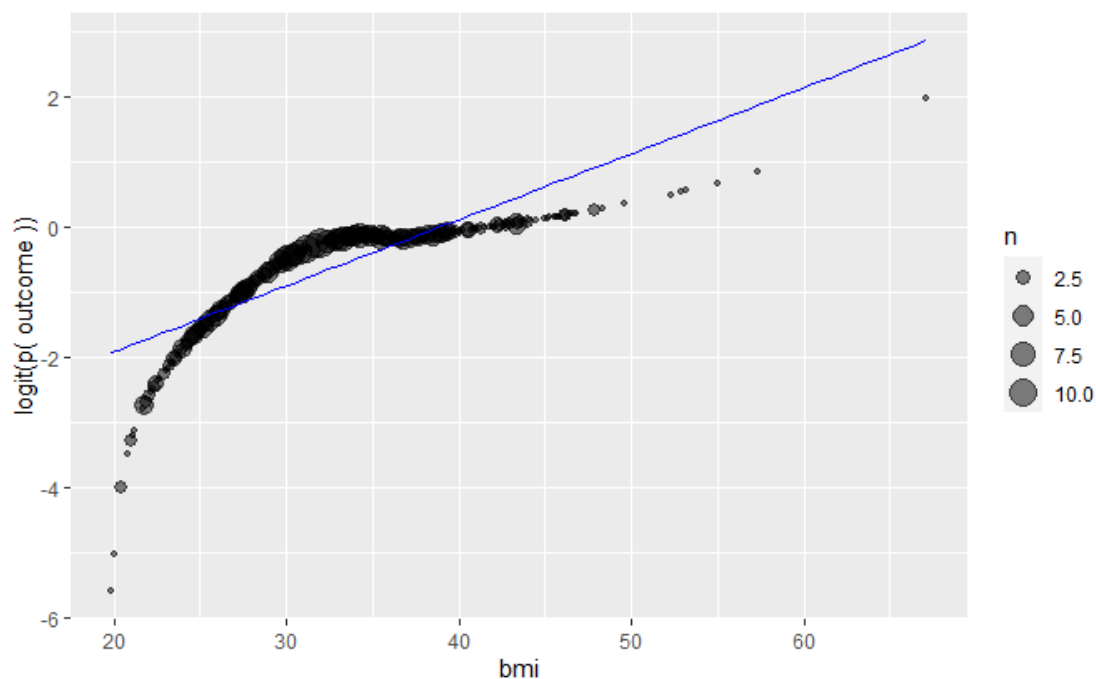
```
## 
## Call:
## glm(formula = outcome ~ ins, family = binomial, data = db_train)
## 
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3307  -0.8481  -0.7291   1.2979   1.7760
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.507593   0.235335  -6.406 1.49e-10 ***
## ins          0.005585   0.001210   4.616 3.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 365.87  on 280  degrees of freedom
## Residual deviance: 339.72  on 279  degrees of freedom
##   (268 observations deleted due to missingness)
## AIC: 343.72
## 
## Number of Fisher Scoring iterations: 4
## 
##                   2.5 %       97.5 %
## (Intercept) -1.985428325 -1.060450949
## ins          0.003325601  0.008089575
```

```
## Call:
## mfp::mfp(formula = outcome ~ fp(bmi), data = db_train, family = binomial)
##
##
## Deviance table:
##              Resid. Dev
## Null model     710.4404
## Linear model   664.1256
## Final model    651.2348
##
## Fractional polynomials:
##      df.initial select alpha df.final power1 power2
## bmi           4      1  0.05        2     -2      .
##
##
## Transformations of covariates:
##              formula
## bmi I((bmi/100)^-2)
##
## Rescaled coefficients:
## Intercept       bmi.1
##     1.185   -1708.555
##
## Degrees of Freedom: 541 Total (i.e. Null);  540 Residual
## Null Deviance:        710.4
## Residual Deviance: 651.2      AIC: 655.2
```



```
##
## Call:
```
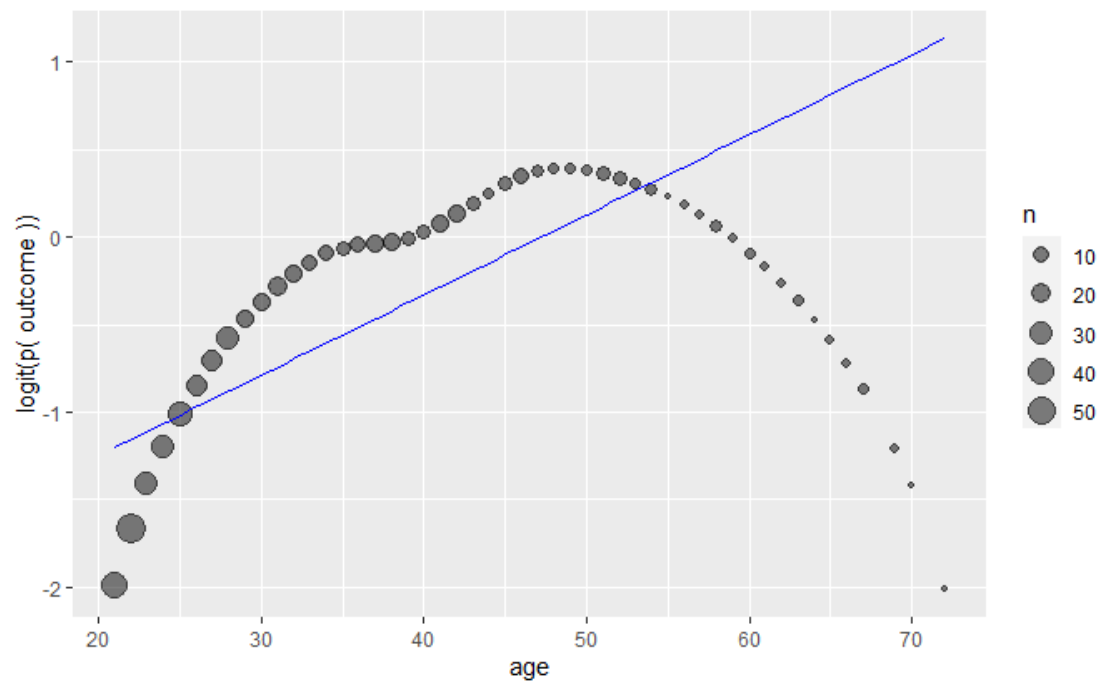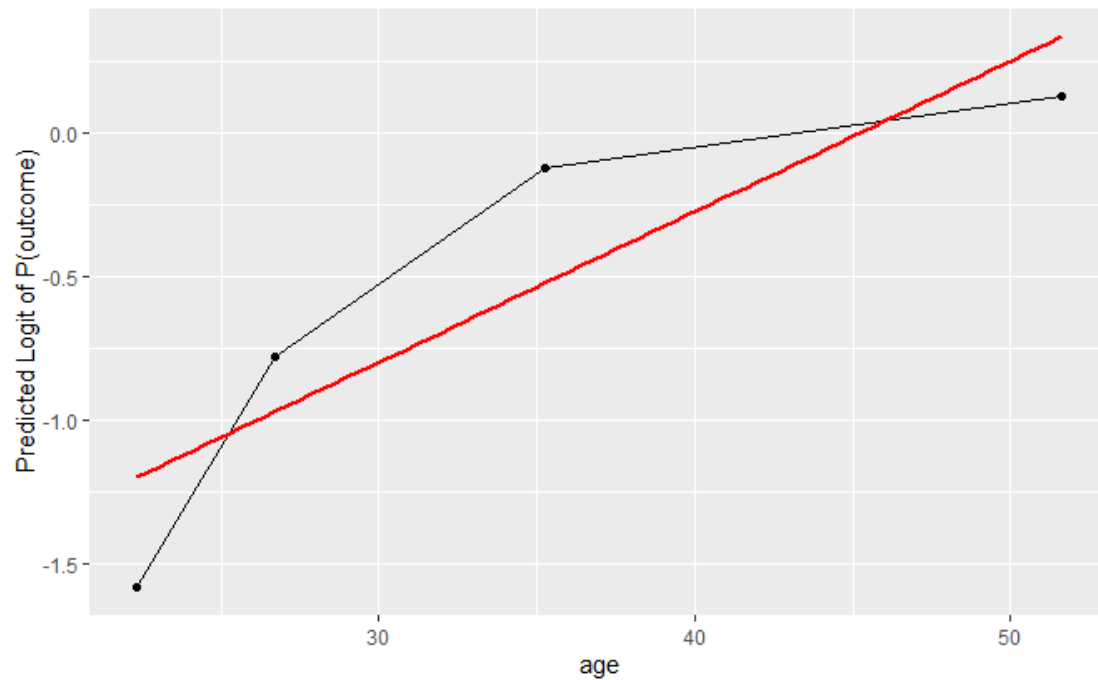
```
## glm(formula = outcome ~ bmi, family = binomial, data = db_train)
##
## Deviance Residuals:
##     Min      1Q   Median      3Q      Max
## -1.9231  -0.9436  -0.7010   1.2318   1.8403
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.59904    0.48970   -7.35 1.99e-13 ***
## bmi          0.09209    0.01441    6.39 1.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 710.44  on 541  degrees of freedom
## Residual deviance: 664.13  on 540  degrees of freedom
##   (7 observations deleted due to missingness)
## AIC: 668.13
##
## Number of Fisher Scoring iterations: 4

##                    2.5 %      97.5 %
## (Intercept) -4.58442790 -2.6621860
## bmi          0.06446343  0.1210361

## Call:
## mfp::mfp(formula = outcome ~ age, data = db_train, family = binomial)
##
##
## Deviance table:
##             Resid. Dev
## Null model    719.0036
## Linear model  691.5872
## Final model   691.5872
##
## Fractional polynomials:
##     df.initial select alpha df.final power1 power2
## age          1      1  0.05        1      1      .
##
##
## Transformations of covariates:
##     formula
## age     age
##
## Rescaled coefficients:
## Intercept      age.1
##  -1.89283    0.03929
##
## Degrees of Freedom: 548 Total (i.e. Null);  547 Residual
```

```
## Null Deviance:          719
## Residual Deviance: 691.6      AIC: 695.6

## Analysis of Deviance Table
##
## Model 1: y ~ meanx
## Model 2: y ~ factor(meanx)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       547     678.78
## 2       545     667.31  2   11.471 0.003229 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## glm(formula = outcome ~ age, family = binomial, data = db_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7508  -0.8783  -0.7819   1.2821   1.6512
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.892829   0.277508  -6.821 9.05e-12 ***
## age          0.039286   0.007683   5.113 3.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 719.00  on 548  degrees of freedom
## Residual deviance: 691.59  on 547  degrees of freedom
## AIC: 695.59
##
## Number of Fisher Scoring iterations: 4

##                  2.5 %       97.5 %
## (Intercept) -2.44575024 -1.35649365
## age          0.02441305  0.05458893
```

Preliminary main effect model

Interaction Term and Counfounder

```
##
## Call:
## glm(formula = outcome ~ dpf + lip + glu + ins + bmi + ins * dpf,
##     family = binomial, data = db_train)
##
## Deviance Residuals:
##     Min      1Q   Median       3Q      Max
## -2.7075  -0.6115  -0.3017   0.6090   2.7659
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.211794   1.592786  -7.039 1.93e-12 ***
## dpf           3.233251   0.942372   3.431 0.000601 ***
## lip           0.129895   0.057029   2.278 0.022743 *
## glu           0.045153   0.007833   5.764 8.20e-09 ***
## ins           0.005033   0.002535   1.986 0.047054 *
## bmi           0.078856   0.030277   2.604 0.009202 **
## dpf:ins      -0.008809   0.003100  -2.842 0.004484 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 310.36  on 240  degrees of freedom
## Residual deviance: 201.51  on 234  degrees of freedom
##   (308 observations deleted due to missingness)
## AIC: 215.51
##
## Number of Fisher Scoring iterations: 5
```

```
## 
## Call:
## glm(formula = outcome ~ dpf + lip + glu + ins + bmi + age + ins *
##     dpf, family = binomial, data = db_train)
## 
## Deviance Residuals:
##     Min      1Q   Median      3Q      Max
## -2.4441  -0.6143  -0.2782   0.6201   2.8004
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.153168   1.708079  -7.115 1.12e-12 ***
## dpf           3.035229   0.942101   3.222  0.00127 **
## lip           0.031258   0.076386   0.409  0.68238
## glu           0.043467   0.007872   5.522 3.35e-08 ***
## ins           0.004675   0.002525   1.851  0.06415 .
## bmi           0.085682   0.030519   2.807  0.00499 **
## age           0.043374   0.023047   1.882  0.05984 .
## dpf:ins      -0.008254   0.003028  -2.726  0.00641 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 310.36  on 240  degrees of freedom
## Residual deviance: 197.80  on 233  degrees of freedom
##   (308 observations deleted due to missingness)
## AIC: 213.8
## 
## Number of Fisher Scoring iterations: 5
```

- The parameter estimate for lipotoxicity changed by 68% after including age into the model. It appears that age confounds the relationship between lipotoxicity and diabetes outcome.

```
## 
## Call:
## glm(formula = outcome ~ lip, family = binomial, data = db_train)
## 
## Deviance Residuals:
##     Min      1Q   Median      3Q      Max
## -1.5271  -0.9061  -0.7358   1.2582   1.6966
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.33192    0.18038  -7.384 1.53e-13 ***
## lip          0.16343    0.03169   5.157 2.50e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 616.92  on 471  degrees of freedom
## Residual deviance: 588.93  on 470  degrees of freedom
##   (77 observations deleted due to missingness)
## AIC: 592.93
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = outcome ~ lip + age, family = binomial, data = db_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7599  -0.8963  -0.7119   1.1773   1.7879
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.135829   0.312415  -6.837 8.11e-12 ***
## lip          0.097365   0.037316   2.609  0.00907 **
## age          0.031718   0.009803   3.236  0.00121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 616.92  on 471  degrees of freedom
## Residual deviance: 578.30  on 469  degrees of freedom
##   (77 observations deleted due to missingness)
## AIC: 584.3
##
## Number of Fisher Scoring iterations: 4
```

Final Model Statistics

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  final_model$y, fitted(final_model)
## X-squared = 8.125, df = 18, p-value = 0.9767
```

Outlier and Leverage Diagnostics for outcome
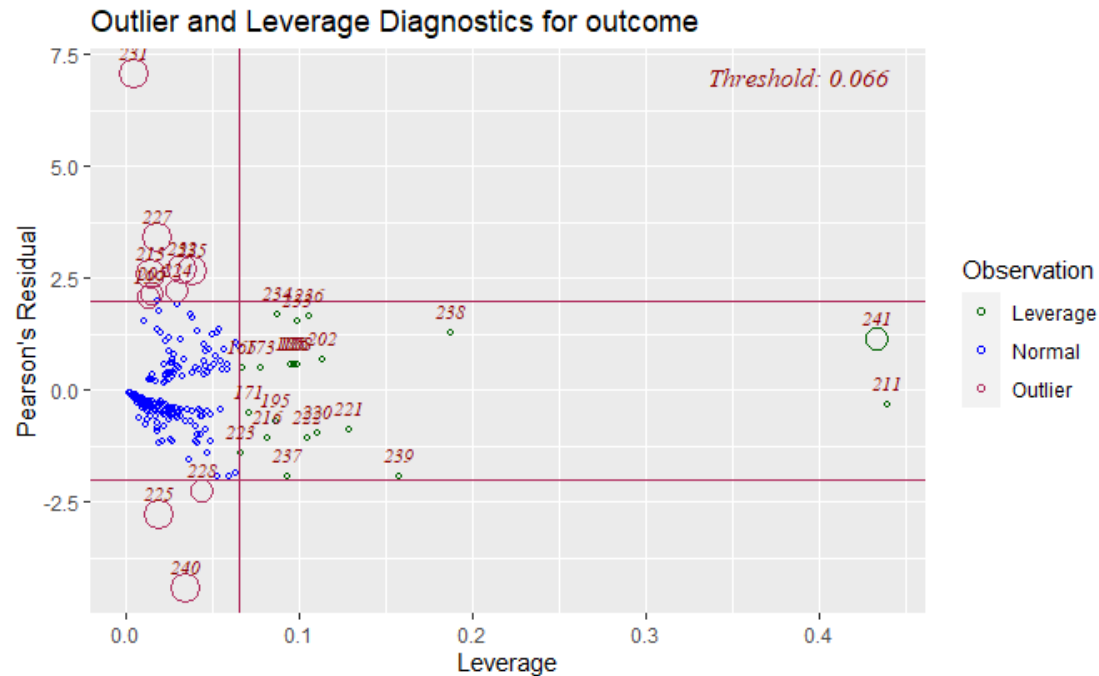
```
## 
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   1   0
##          1  55  20
##          0  28 138
## 
##                 Total n : 241
##                Accuracy : 0.8008
##                  95% CI : (0.7459, 0.8464)
##     No Information Rate : 0.6556
##     P-Value [Acc > NIR] : 5.26e-07
## 
##                   Kappa : 0.5486
##  Mcnemar's Test P-Value : 0.3123
## 
##             Sensitivity : 0.6627
##             Specificity : 0.8734
##          Pos Pred Value : 0.7333
##          Neg Pred Value : 0.8313
##              Prevalence : 0.3444
##          Detection Rate : 0.3112
##    Detection Prevalence : 0.2282
##       Balanced Accuracy : 0.7680
##           F-val Accuracy : 0.6962
##      Matthews Cor.-Coef : 0.5502
## 
##        'Positive' Class : 1
```
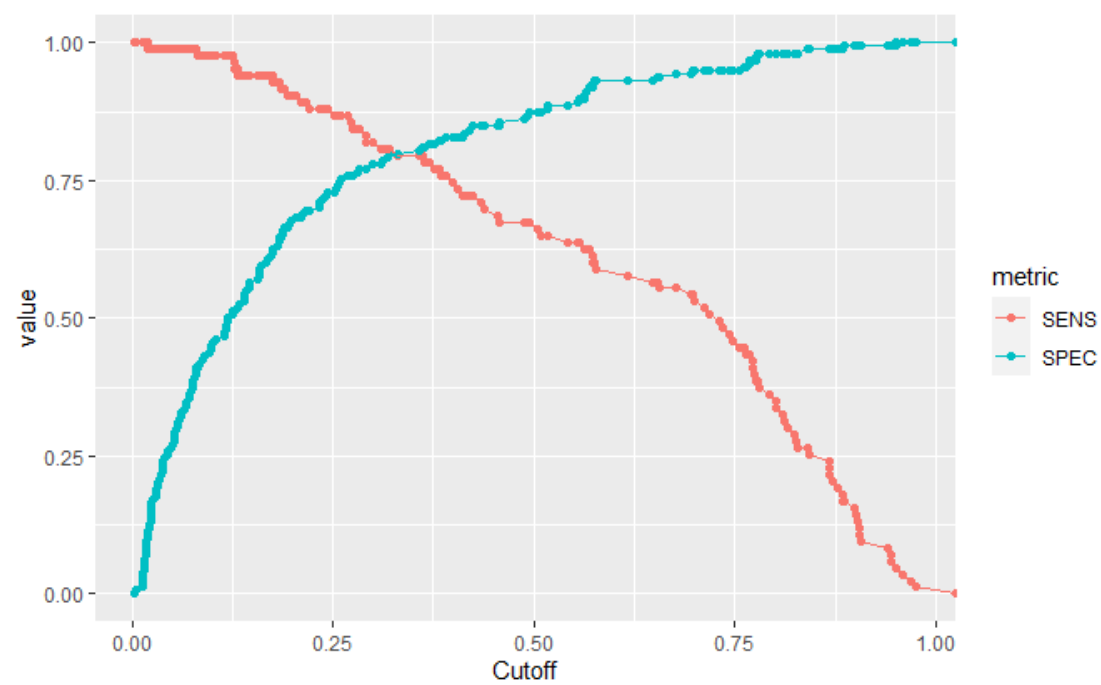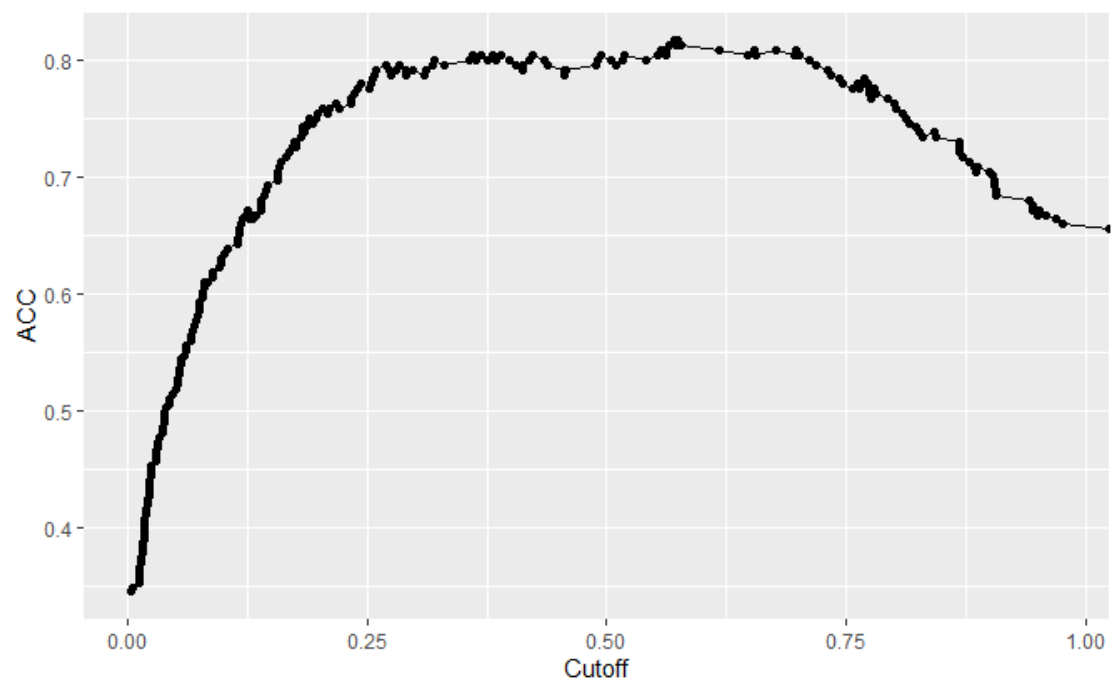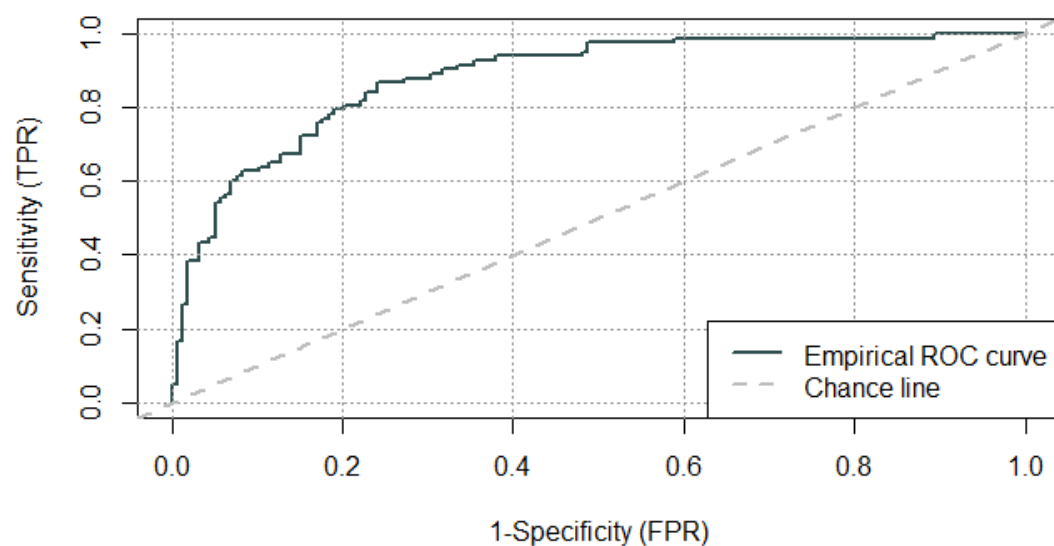
```
##
##  Method used: empirical
##  Number of positive(s): 83
##  Number of negative(s): 158
##  Area under curve: 0.8817

##
##    estimated AUC : 0.881729449443343
##    AUC estimation method : empirical
##
##    CI of AUC
##    confidence level = 95%
##    lower = 0.831304105800821      upper = 0.932154793085865

##
## Call:
## optimal.cutpoints.default(X = "pred_p", status = "y", tag.healthy = 0,
##     methods = c("Youden", "MaxSpSe", "MaxProdSpSe"), data = data.frame(fin
al_model.p))
##
## Optimal cutoffs:
##    Youden MaxSpSe MaxProdSpSe
## 1 0.2692  0.3202       0.2692
##
## Area under the ROC curve (AUC):  0.882 (0.838, 0.925)
```

Final Model Applying Test Dataset

```
##
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction  1   0
##          1 16   7
##          0 12  60
##
##                Total n : 95
##               Accuracy : 0.8000
##                 95% CI : (0.7086, 0.8681)
##    No Information Rate : 0.7053
##    P-Value [Acc > NIR] : 0.0248
##
##                  Kappa : 0.4925
##  Mcnemar's Test P-Value : 0.3588
##
##            Sensitivity : 0.5714
##            Specificity : 0.8955
##         Pos Pred Value : 0.6957
##         Neg Pred Value : 0.8333
##             Prevalence : 0.2947
##         Detection Rate : 0.2421
##   Detection Prevalence : 0.1684
##       Balanced Accuracy : 0.7335
##         F-val Accuracy : 0.6275
##     Matthews Cor.-Coef : 0.4970
##
##        'Positive' Class : 1
```

Data summary

Name	test_model.p
Number of rows	219
Number of columns	2

_____

Column type frequency:	
numeric	2

_____

| Group variables | None |

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
pred_p	124	0.43	0.36	0.30	0	0.1	0.29	0.51	0.99	▆▁▂▁▃

y   0   1.00   0.32   0.4   0   0.0   0.0   1.0   1.00   ▮▁▁▁
                                              7   0   0   ▮
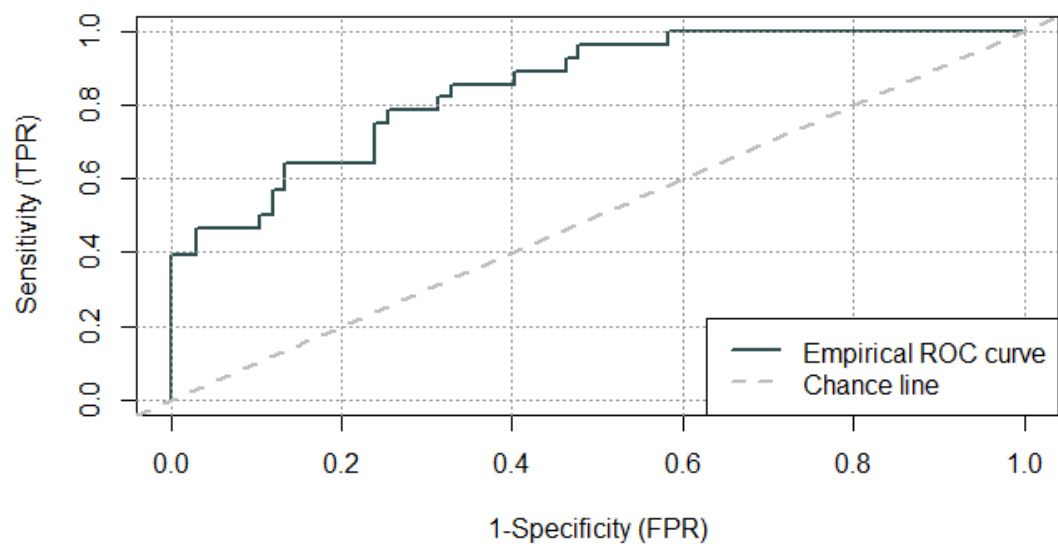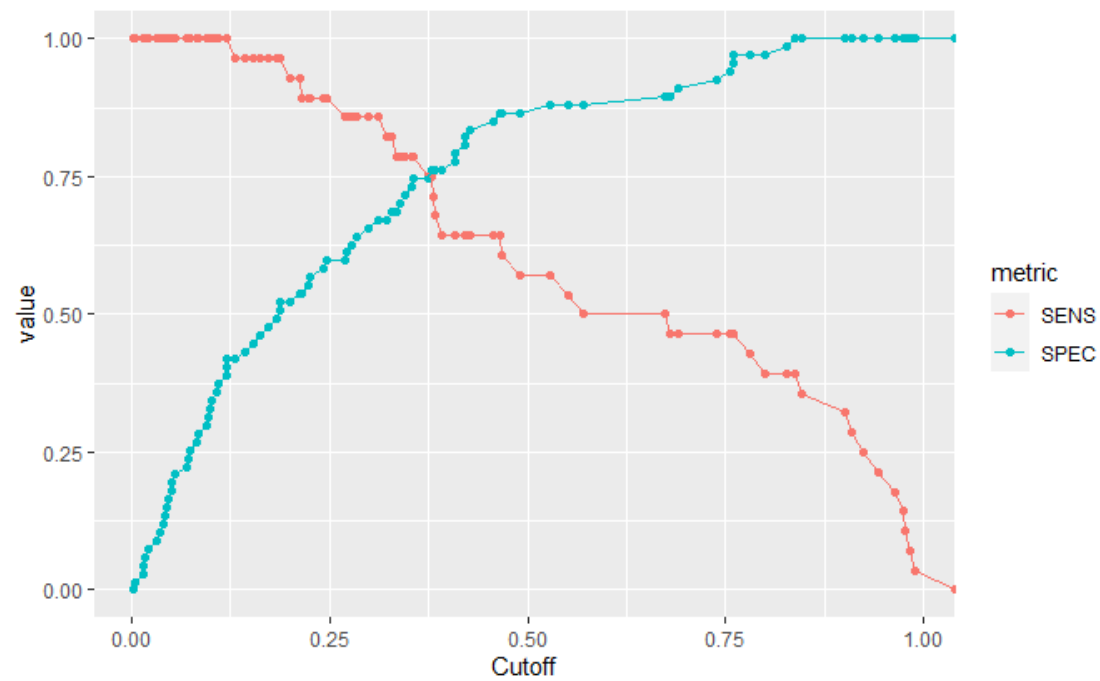

```
## 
##  Method used: empirical
##  Number of positive(s): 28
##  Number of negative(s): 67
##  Area under curve: 0.8497
```

```
## 
##     estimated AUC : 0.849680170575693
##     AUC estimation method : empirical
## 
##     CI of AUC
##     confidence level = 95%
##     lower = 0.754214626539895      upper = 0.945145714611491
```