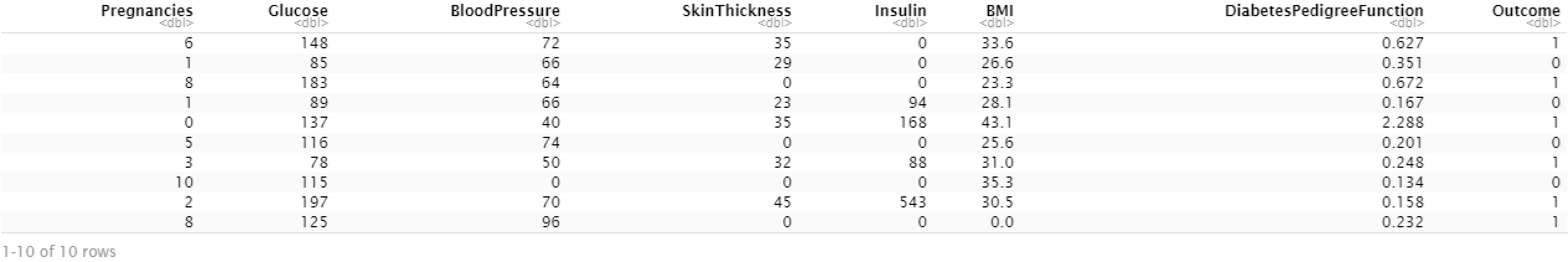
Analyzing Diabetes Risk Factors

Jina Park

Background Information

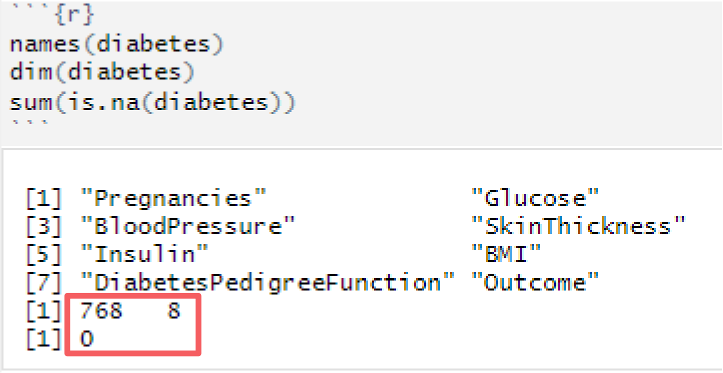
- Number of people with diabetes is **increasing** every year
- Major causes of **blindness, kidney failure, heart attacks, stroke, and lower limb amputation**
- Constraints : **females** of **age 21 and above** & **Pima Indian heritage**

Data



Pregnancies <dbl>	Glucose <dbl>	BloodPressure <dbl>	SkinThickness <dbl>	Insulin <dbl>	BMI <dbl>	DiabetesPedigreeFunction <dbl>	Outcome <dbl>
6	148	72	35	0	33.6	0.627	1
1	85	66	29	0	26.6	0.351	0
8	183	64	0	0	23.3	0.672	1
1	89	66	23	94	28.1	0.167	0
0	137	40	35	168	43.1	2.288	1
5	116	74	0	0	25.6	0.201	0
3	78	50	32	88	31.0	0.248	1
10	115	0	0	0	35.3	0.134	0
2	197	70	45	543	30.5	0.158	1
8	125	96	0	0	0.0	0.232	1

1-10 of 10 rows



~~~
```{r}
names(diabetes)
dim(diabetes)
sum(is.na(diabetes))
```
~~~

```
 [1] "Pregnancies"              "Glucose"
 [3] "BloodPressure"            "SkinThickness"
 [5] "Insulin"                  "BMI"
 [7] "DiabetesPedigreeFunction" "Outcome"
[1] 768   8
[1] 0
```

8 variables & 768 individual data

No missing data

Variables and Descriptions

Pregnancies: Number of pregnancies

Glucose: Glucose concentration

Blood Pressure: Blood pressure in mm/Hg

Skin Thickness: Triceps skinfold thickness in mm

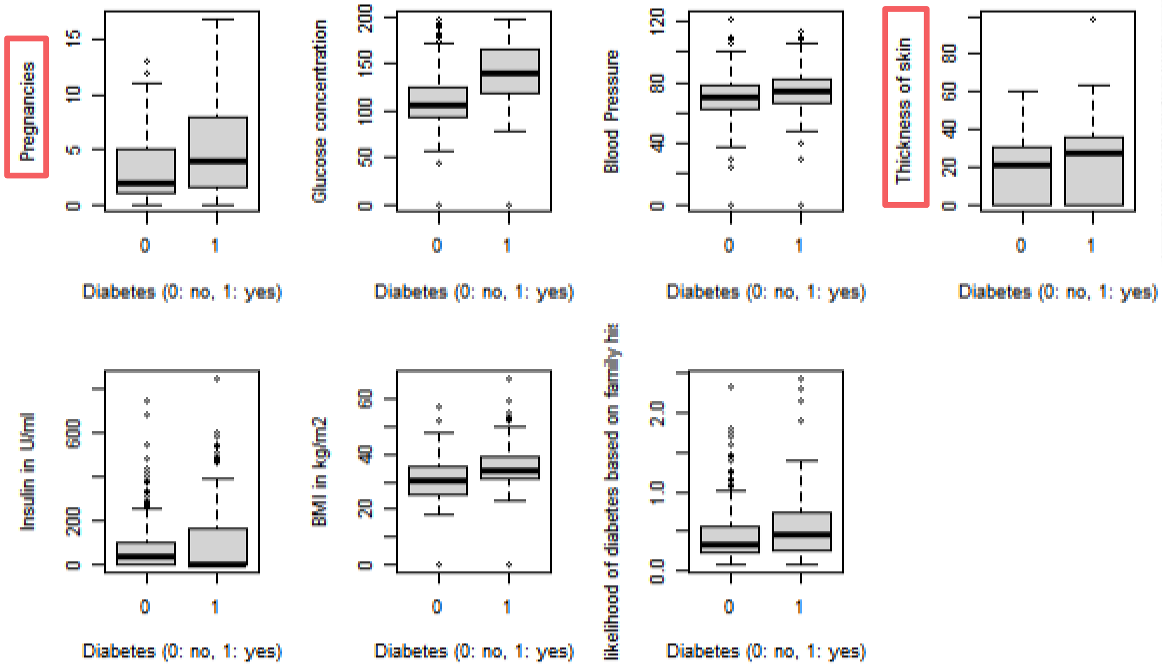
Insulin: Insulin in U/mL

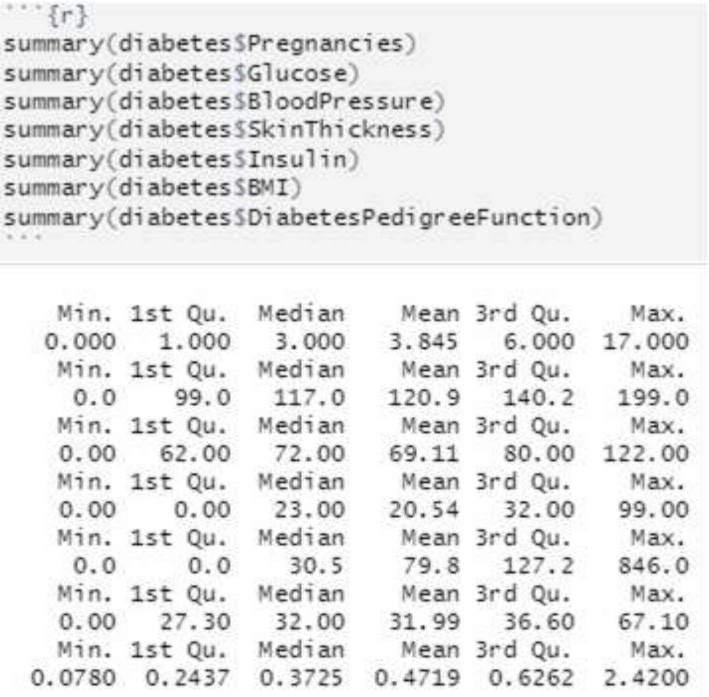
BMI: Body mass index in kg/m2

Diabetes Pedigree Function: function that scores likelihood of diabetes based on family history

Outcome: 1 = have diabetes, 0 = no diabetes

Boxplots

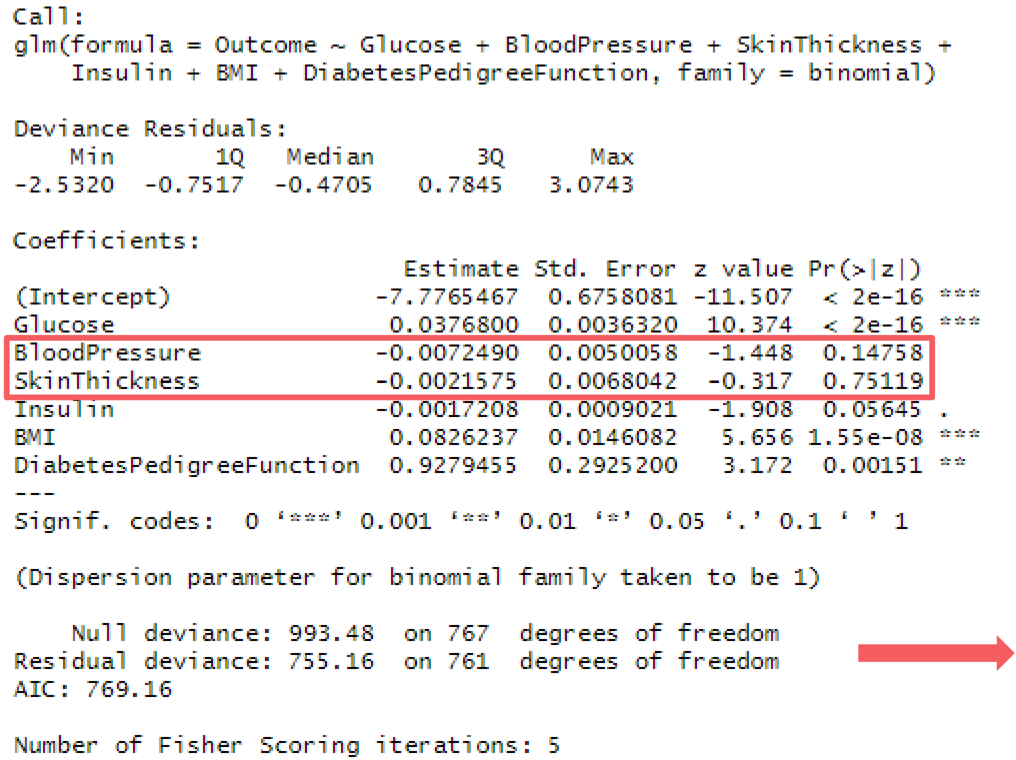




~~~
```{r}
summary(diabetes$Pregnancies)
summary(diabetes$Glucose)
summary(diabetes$BloodPressure)
summary(diabetes$SkinThickness)
summary(diabetes$Insulin)
summary(diabetes$BMI)
summary(diabetes$DiabetesPedigreeFunction)
```
~~~

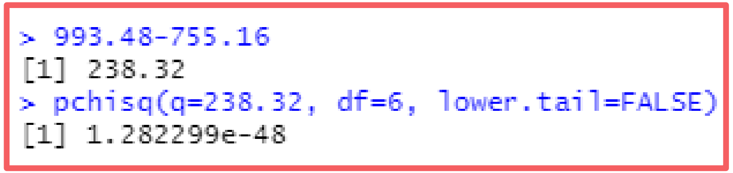
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.000   3.000   3.845   6.000  17.000 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0    99.0   117.0   120.9   140.2   199.0 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   62.00   72.00   69.11   80.00  122.00 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.00   23.00   20.54   32.00   99.00 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0     0.0    30.5    79.8   127.2   846.0 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   27.30   32.00   31.99   36.60   67.10 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0780  0.2437  0.3725  0.4719  0.6262  2.4200 
```

Model 1



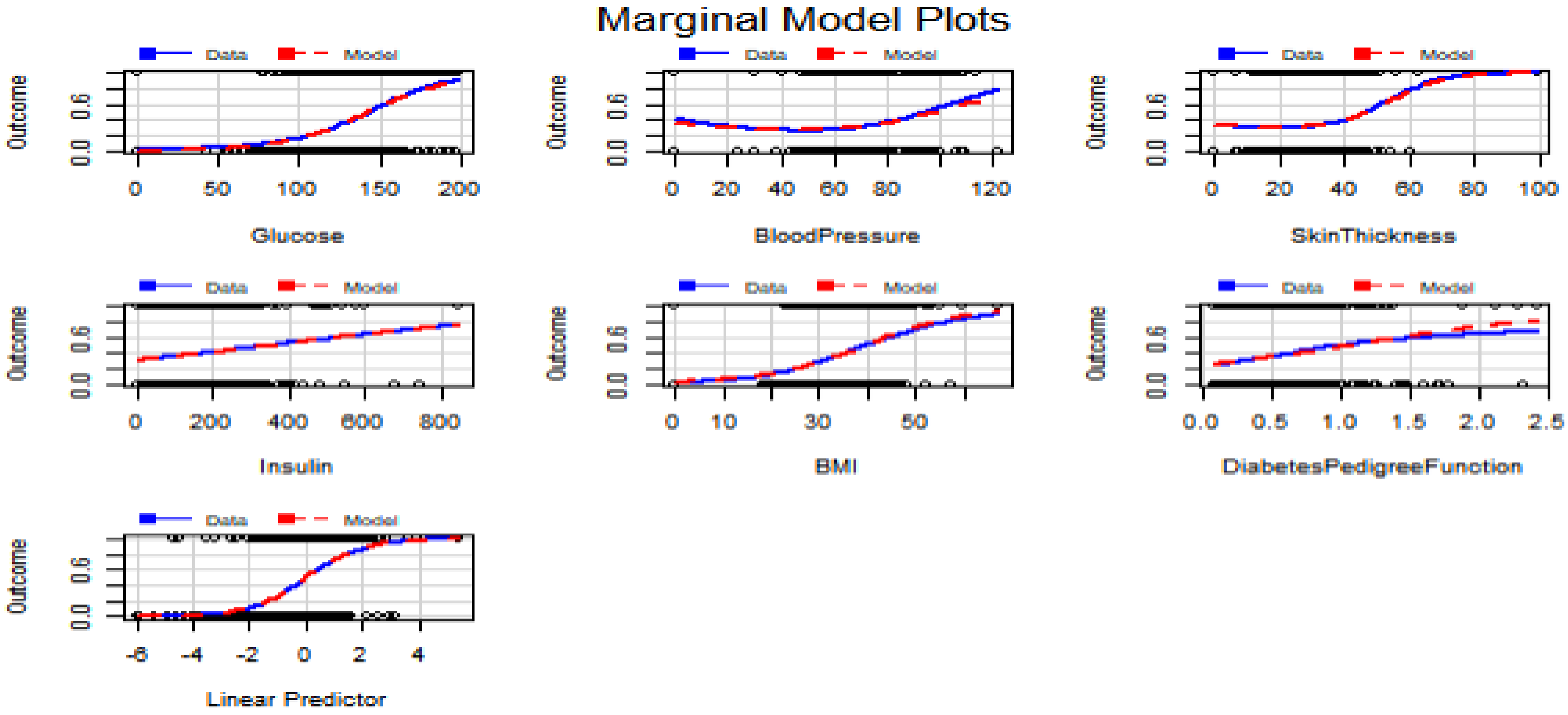
```
Call:
glm(formula = Outcome ~ Glucose + BloodPressure + SkinThickness + 
    Insulin + BMI + DiabetesPedigreeFunction, family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5320  -0.7517  -0.4705   0.7845   3.0743  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -7.7765467  0.6758081 -11.507  < 2e-16 ***
Glucose                   0.0376800  0.0036320  10.374  < 2e-16 ***
BloodPressure            -0.0072490  0.0050058  -1.448  0.14758    
SkinThickness            -0.0021575  0.0068042  -0.317  0.75119    
Insulin                  -0.0017208  0.0009021  -1.908  0.05645 .  
BMI                       0.0826237  0.0146082   5.656 1.55e-08 ***
DiabetesPedigreeFunction  0.9279455  0.2925200   3.172  0.00151 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 755.16  on 761  degrees of freedom
AIC: 769.16

Number of Fisher Scoring iterations: 5
```

- Blood pressure and Skin thickness : Insignificant



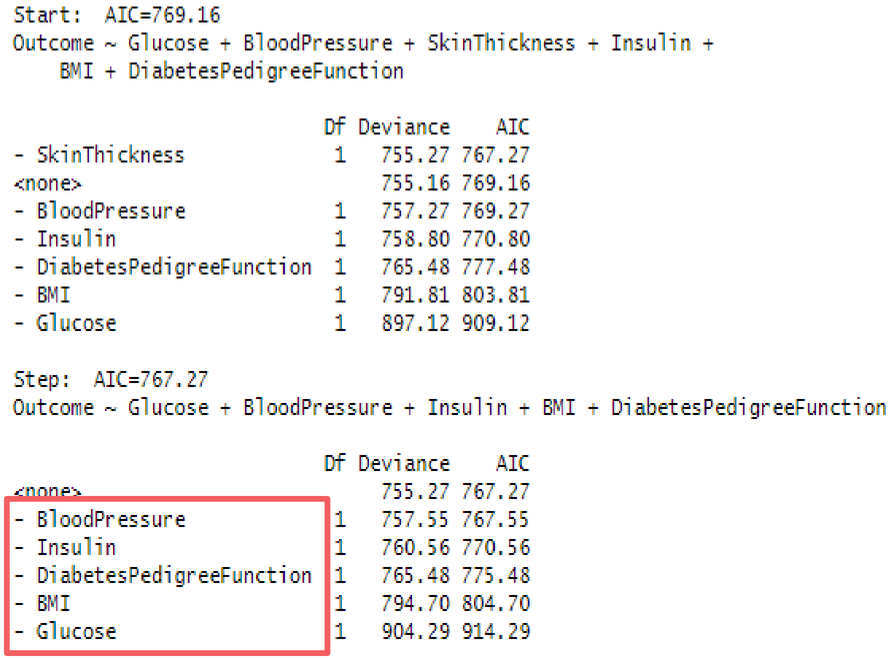
```
> 993.48-755.16
[1] 238.32
> pchisq(q=238.32, df=6, lower.tail=FALSE)
[1] 1.282299e-48
```

Model 1 - MMPs


Marginal Model Plots
Data
Model
Outcome
0.0
0.6
0
50
100
150
200
Glucose
0
20
40
60
80
120
BloodPressure
0
20
40
60
80
100
SkinThickness
0
200
400
600
800
Insulin
0
10
30
50
BMI
0.0
0.5
1.0
1.5
2.0
2.5
DiabetesPedigreeFunction
-6
-4
-2
0
2
4
Linear Predictor

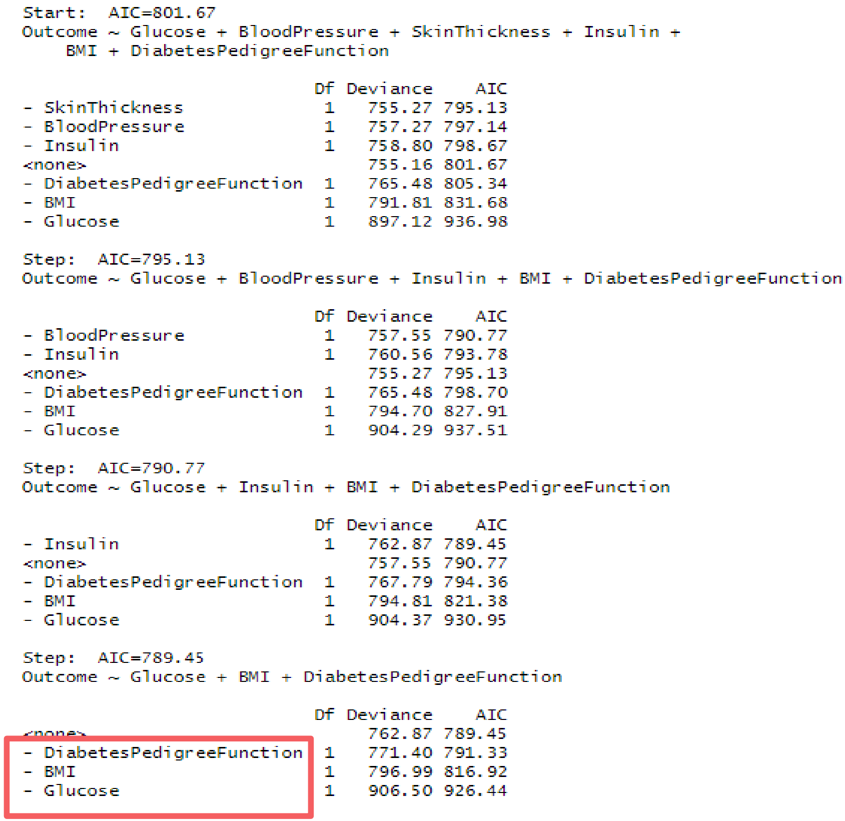
AIC and BIC methods

→ AIC



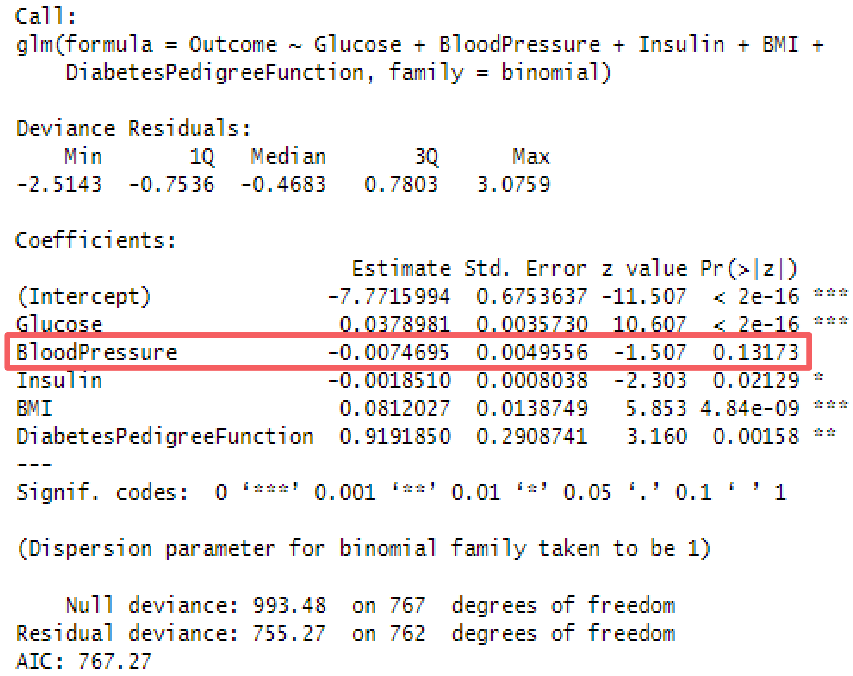
```
Start:  AIC=769.16
Outcome ~ Glucose + BloodPressure + SkinThickness + Insulin + 
    BMI + DiabetesPedigreeFunction

                           Df Deviance    AIC
- SkinThickness             1   755.27 767.27
<none>                          755.16 769.16
- BloodPressure             1   757.27 769.27
- Insulin                   1   758.80 770.80
- DiabetesPedigreeFunction  1   765.48 777.48
- BMI                       1   791.81 803.81
- Glucose                   1   897.12 909.12

Step:  AIC=767.27
Outcome ~ Glucose + BloodPressure + Insulin + BMI + DiabetesPedigreeFunction

                           Df Deviance    AIC
<none>                          755.27 767.27
- BloodPressure             1   757.55 767.55
- Insulin                   1   760.56 770.56
- DiabetesPedigreeFunction  1   765.48 775.48
- BMI                       1   794.70 804.70
- Glucose                   1   904.29 914.29
```

→ BIC

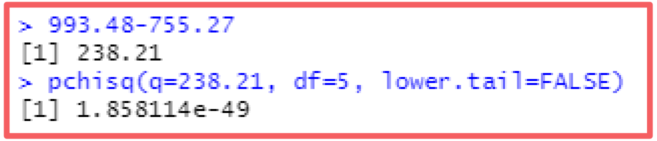


```
Start:  AIC=801.67
Outcome ~ Glucose + BloodPressure + SkinThickness + Insulin + 
    BMI + DiabetesPedigreeFunction

                           Df Deviance    AIC
- SkinThickness             1   755.27 795.13
- BloodPressure             1   757.27 797.14
- Insulin                   1   758.80 798.67
<none>                          755.16 801.67
- DiabetesPedigreeFunction  1   765.48 805.34
- BMI                       1   791.81 831.68
- Glucose                   1   897.12 936.98

Step:  AIC=795.13
Outcome ~ Glucose + BloodPressure + Insulin + BMI + DiabetesPedigreeFunction

                           Df Deviance    AIC
- BloodPressure             1   757.55 790.77
- Insulin                   1   760.56 793.78
<none>                          755.27 795.13
- DiabetesPedigreeFunction  1   765.48 798.70
- BMI                       1   794.70 827.91
- Glucose                   1   904.29 937.51

Step:  AIC=790.77
Outcome ~ Glucose + Insulin + BMI + DiabetesPedigreeFunction

                           Df Deviance    AIC
- Insulin                   1   762.87 789.45
<none>                          757.55 790.77
- DiabetesPedigreeFunction  1   767.79 794.36
- BMI                       1   794.81 821.38
- Glucose                   1   904.37 930.95

Step:  AIC=789.45
Outcome ~ Glucose + BMI + DiabetesPedigreeFunction

                           Df Deviance    AIC
<none>                          762.87 789.45
- DiabetesPedigreeFunction  1   771.40 791.33
- BMI                       1   796.99 816.92
- Glucose                   1   906.50 926.44
```

Model 2 - AIC

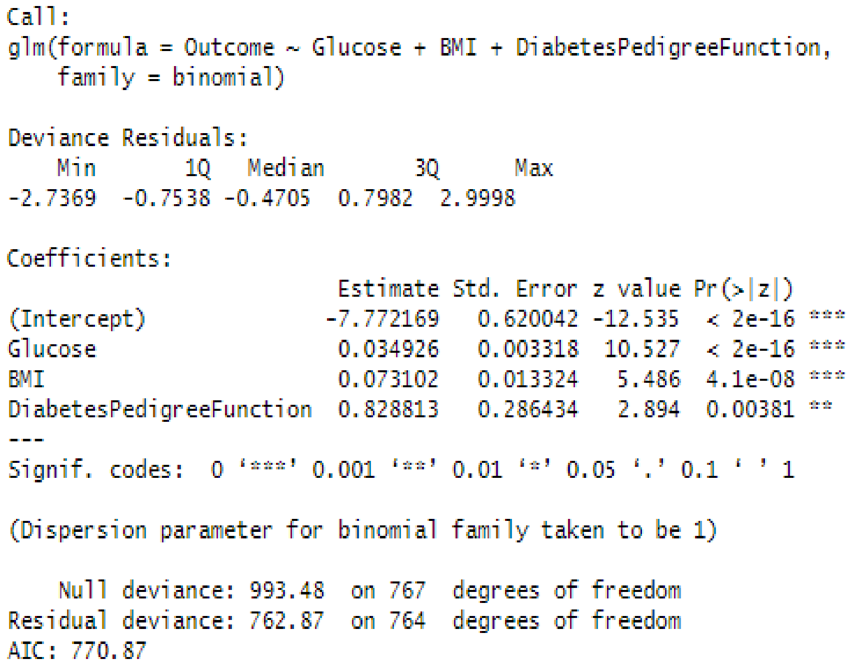


```
Call:
glm(formula = Outcome ~ Glucose + BloodPressure + Insulin + BMI + 
    DiabetesPedigreeFunction, family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5143  -0.7536  -0.4683   0.7803   3.0759  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -7.7715994  0.6753637 -11.507  < 2e-16 ***
Glucose                   0.0378981  0.0035730  10.607  < 2e-16 ***
BloodPressure            -0.0074695  0.0049556  -1.507  0.13173    
Insulin                  -0.0018510  0.0008038  -2.303  0.02129 *  
BMI                       0.0812027  0.0138749   5.853 4.84e-09 ***
DiabetesPedigreeFunction  0.9191850  0.2908741   3.160  0.00158 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 755.27  on 762  degrees of freedom
AIC: 767.27
```

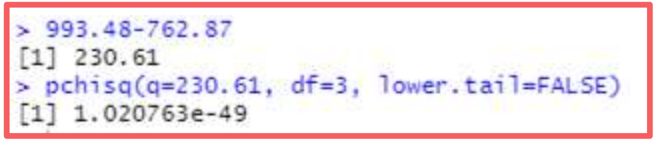


```
> 993.48-755.27
[1] 238.21
> pchisq(q=238.21, df=5, lower.tail=FALSE)
[1] 1.858114e-49
```

Model 3 - BIC

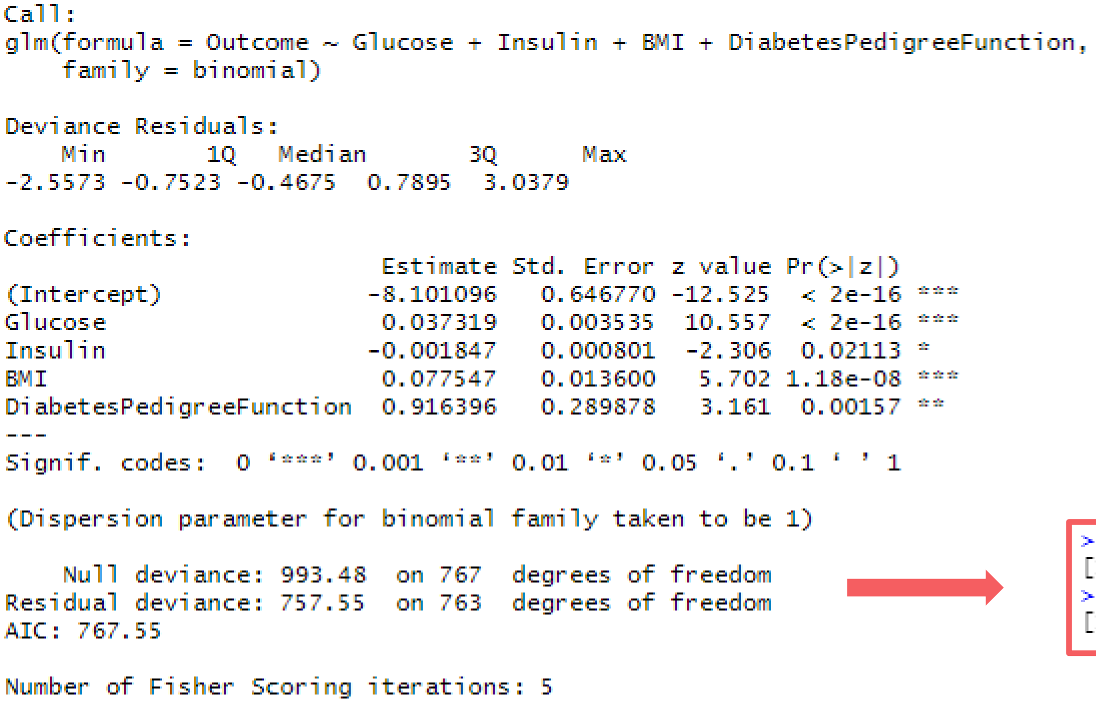


```
Call:
glm(formula = Outcome ~ Glucose + BMI + DiabetesPedigreeFunction, 
    family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7369  -0.7538  -0.4705   0.7982   2.9998  

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -7.772169   0.620042 -12.535  < 2e-16 ***
Glucose                   0.034926   0.003318  10.527  < 2e-16 ***
BMI                       0.073102   0.013324   5.486  4.1e-08 ***
DiabetesPedigreeFunction  0.828813   0.286434   2.894  0.00381 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 762.87  on 764  degrees of freedom
AIC: 770.87
```

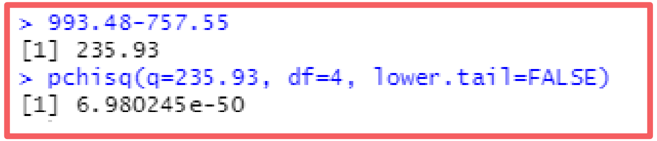


```
> 993.48-762.87
[1] 230.61
> pchisq(q=230.61, df=3, lower.tail=FALSE)
[1] 1.020763e-49
```

Model 4



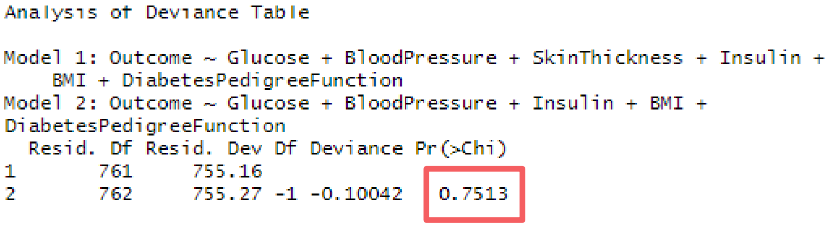
```
Call:
glm(formula = Outcome ~ Glucose + Insulin + BMI + DiabetesPedigreeFunction, 
    family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5573  -0.7523  -0.4675   0.7895   3.0379  

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -8.101096   0.646770 -12.525  < 2e-16 ***
Glucose                   0.037319   0.003535  10.557  < 2e-16 ***
Insulin                  -0.001847   0.000801  -2.306  0.02113 *  
BMI                       0.077547   0.013600   5.702 1.18e-08 ***
DiabetesPedigreeFunction  0.916396   0.289878   3.161  0.00157 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 757.55  on 763  degrees of freedom
AIC: 767.55

Number of Fisher Scoring iterations: 5
```



```
> 993.48-757.55
[1] 235.93
> pchisq(q=235.93, df=4, lower.tail=FALSE)
[1] 6.980245e-50
```

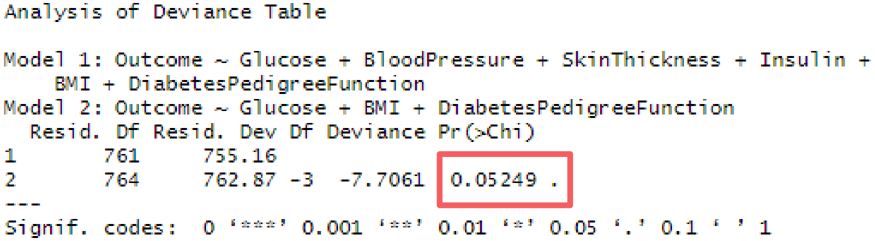
ANOVA Tests

→ Compare models 1 & 2



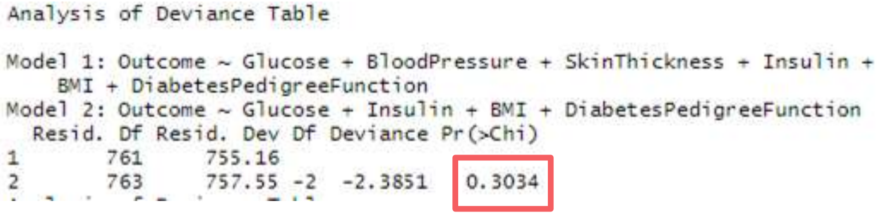
```
Analysis of Deviance Table

Model 1: Outcome ~ Glucose + BloodPressure + SkinThickness + Insulin + 
    BMI + DiabetesPedigreeFunction
Model 2: Outcome ~ Glucose + BloodPressure + Insulin + BMI + 
DiabetesPedigreeFunction
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       761     755.16                     
2       762     755.27 -1 -0.10042   0.7513
```

→ Compare models 1 & 3



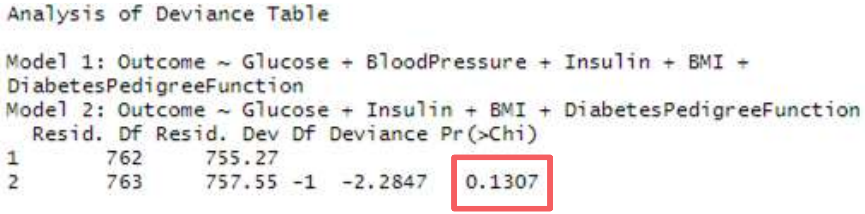
```
Analysis of Deviance Table

Model 1: Outcome ~ Glucose + BloodPressure + SkinThickness + Insulin + 
    BMI + DiabetesPedigreeFunction
Model 2: Outcome ~ Glucose + BMI + DiabetesPedigreeFunction
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
1       761     755.16                       
2       764     762.87 -3  -7.7061  0.05249 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

→ Compare models 1& 4



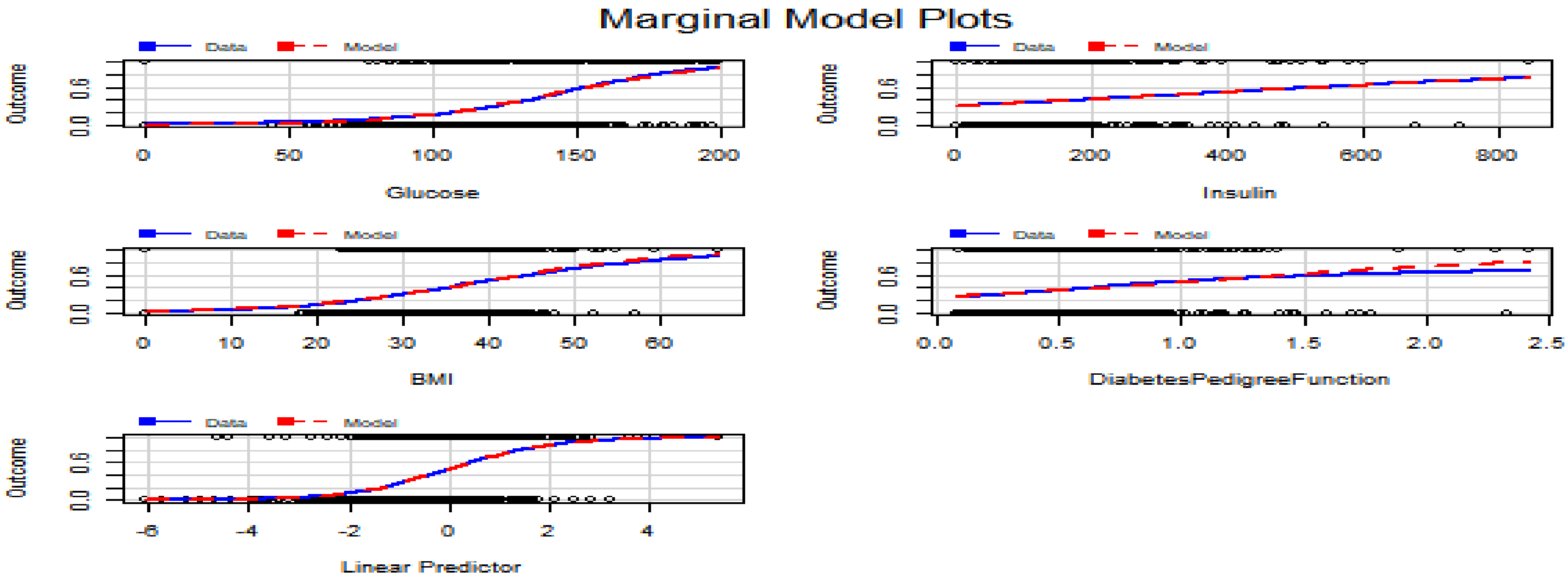
```
Analysis of Deviance Table

Model 1: Outcome ~ Glucose + BloodPressure + SkinThickness + Insulin + 
    BMI + DiabetesPedigreeFunction
Model 2: Outcome ~ Glucose + Insulin + BMI + DiabetesPedigreeFunction
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       761     755.16                     
2       763     757.55 -2  -2.3851   0.3034
```

→ Compare models 2& 4



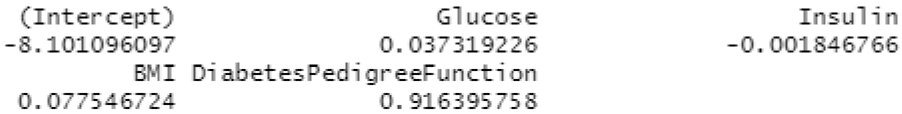
```
Analysis of Deviance Table

Model 1: Outcome ~ Glucose + BloodPressure + Insulin + BMI + 
DiabetesPedigreeFunction
Model 2: Outcome ~ Glucose + Insulin + BMI + DiabetesPedigreeFunction
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       762     755.27                     
2       763     757.55 -1  -2.2847   0.1307
```

Model 4 - MMPs and Multicollinearity



```
                 Glucose                  Insulin                      BMI
                1.126073                 1.174711                 1.030206
DiabetesPedigreeFunction
                1.023711
```

Conclusion



```
     (Intercept)                  Glucose                  Insulin
    -8.101096097              0.037319226             -0.001846766
             BMI DiabetesPedigreeFunction
     0.077546724              0.916395758
```

$$P(Diabetes) = \frac{exp(-8.101 + 0.037Glucose - 0.002Insulin + 0.076BMI + 0.916DiabetesPedigreeFunction)}{1 + exp(-8.101 + 0.037Glucose - 0.002Insulin + 0.076BMI + 0.916DiabetesPedigreeFunction)}$$

- Model 4 is the best model
- Family history > BMI > Glucose concentration > Insulin concentration
- Negative correlation with insulin

Resources

https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

https://www.who.int/news-room/fact-sheets/detail/diabetes