

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/309679945>

# The Study of Pima Indian Diabetes

**Presentation** · October 2016

DOI: 10.13140/RG.2.2.14751.76960

---

CITATIONS

0

---

READS

165

**1 author:**



**Jian Sun**

George Washington University

**16** PUBLICATIONS **0** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Survival Analysis of German Breast Cancer [View project](#)



Analyzing Default Payments of Credit Card Clients in Taiwan [View project](#)

# The Study of Pima Indian Diabetes

## Summary

This paper focus on Pima Indians Diabetes. The research people is Pima Indian Female, which are 768. The original dataset was attached in Appendix 2. Firstly, the 700 observations were randomly picked up from original dataset. Secondly, the goal of this research are Predict the probability that individual females have diabetes in GLM and Detect subgroups of characteristics that are at higher risk of diabetes in Forward and Backward Stepwise Selection. The higher risk subgroups contain 4 variables, PRG, PLASMA, BODY and PEDIGREE. Finally, we give out a tip to prevent diabetes.

## 1 Introduction

Diabetes is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. To study the reason that leading to diabetes, a cluster of dataset about Pima Indian Diabetes was collected. It is consisted of 8 predict variables and 1 response variable. The variables are PRG, PLASMA, BP, THICK, INSULIN, BODY, PEDIGREE, AGE. After randomly selecting 700 observations from 768 patients, 9 variables were taken to fit a generalize linear model to predict the probability that individual females have diabetes. Then, using stepwise selection provided subgroups of characteristics with higher risk of diabetes.

## 2 Data Analysis and Interpretation

### 2.1 Data Explanation

Variables	Explanation
PRG:	Number of times pregnant
PLASMA:	Plasma glucose concentration in saliva
BP:	Diastolic blood pressure
THICK:	Triceps skin fold thickness
INSULIN:	Two Hours serum insulin
BODY:	Body mass index (Weight/Height)
PEDIGREE:	Diabetes pedigree function
AGE:	In years
RESPONSE:	1: Diabetes, 0: Not

For more details, please check Table 1 in Appendix 1.

### 2.2 Data Selection

Randomly picking up 700 total observations from 768 patients was finished in R. The new dataset was named after PID.

The R Code to do this was attached as Table 2 in Appendix 1.

### 2.3 Predicting the probability to suffer from diabetes for individual females

Generalize linear model contributed to solve this problem. In the glm, the binomial family was chosen,

since the response variable was only consisted of 1 and 0.

The generated model is:

$$\text{REP} = -0.8532997 + \text{PRG} * 0.0211334 + \text{PLASMA} * 0.0059682 - \text{BP} * 0.0021338 + \text{THICK} * 0.0005073 - \text{INSULIN} * 0.0002122 + \text{BODY} * 0.0129676 + \text{PEDIGREE} * 0.1361092 + \text{AGE} * 0.0024004$$

Here, doing Cross-Validation checked Generalized Linear Model. The MSE (mean square of error) of this model was 3.322.

What's more, according to the Table 3, 5 variables (PRG, PLASMA, BP, BODY and PEDIGREE) had an important influence on REP, since their P-Value of coefficient < 0.05.

Hence, the adjusted model was,

$$\text{REP} = -0.853 + \text{PRG} * 0.0211 + \text{PLASMA} * 0.00597 - \text{BP} * 0.00213 + \text{BODY} * 0.0130 + \text{PEDIGREE} * 0.136$$

The next, doing Cross-Validation checked Generalized Linear Model. The MSE (mean square of error) of this model was 2.952.

The MSE of adjusted model, 2.952, is less than that of whole model, 3.322.

Finally, the chosen model to predict the probability was:

$$\text{REP} = -0.853 + \text{PRG} * 0.0211 + \text{PLASMA} * 0.00597 - \text{BP} * 0.00213 + \text{BODY} * 0.0130 + \text{PEDIGREE} * 0.136$$

The value of REP was taken as index to return the probability of suffering from diabetes for individual females.

When the value of REP is close to 1, it means the individual female has higher probability to have diabetes.

When the value of REP is close to 0, it means the individual female has a healthy physical body.

The details were attached in Table 3.

## 2.4 Detecting subgroups of characteristics that are at higher risk of diabetes.

For the subgroup of characteristics, forward and backward stepwise selection (in AIC method) were used.

AIC (Akaike information criterion): a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection.

For forward stepwise selection, the result with the least AIC = 688.06 was

$$\text{REP} = -8.0308 + 0.1481 * \text{PRG} + 0.0356 * \text{PLASMA} + 0.0861 * \text{BODY} + 0.9016 * \text{PEDIGREE} - 0.0107 * \text{BP} - 0.00132 * \text{INSULIN}$$

The details result was shown in Table 4. In the result, the P-value of INSULIN = 0.11091 > 0.05. Let the null hypothesis be the coefficient of INSULIN = 0, so the null hypothesis cannot be rejected since 0.11091 > 0.05. Hence the coefficient of INSULIN = 0.

The final model was

$$\text{REP} = -8.0308 + 0.148 * \text{PRG} + 0.0356 * \text{PLASMA} + 0.0861 * \text{BODY} + 0.9016 * \text{PEDIGREE} - 0.0107 * \text{BP}$$

The MSE (mean square of error) of forward stepwise selection was 3.21068.

For backward stepwise selection, the result with the least AIC = 727.67 was

$$\text{REP} = -8.0308 + 0.1481 \cdot \text{PRG} + 0.0356 \cdot \text{PLASMA} + 0.0861 \cdot \text{BODY} + 0.9016 \cdot \text{PEDIGREE} - 0.0107 \cdot \text{BP} - 0.00132 \cdot \text{INSULIN}$$

The details result was shown in Table 5. In the result, the P-value of INSULIN = 0.11091 > 0.05. Let the null hypothesis be the coefficient of INSULIN = 0, so the null hypothesis cannot be rejected since 0.11091 > 0.05. Hence the coefficient of INSULIN = 0.

The final model was

$$\text{REP} = -8.0308 + 0.148 \cdot \text{PRG} + 0.0356 \cdot \text{PLASMA} + 0.0861 \cdot \text{BODY} + 0.9016 \cdot \text{PEDIGREE} - 0.0107 \cdot \text{BP}$$

The MSE (mean square of error) of forward stepwise selection was 3.21068.

The forward stepwise selection method shared the same answer with backward stepwise selection method.

And the coefficient of BP < 0, so BP will not be considered as dangerous variable to cause diabetes among the given variables.

The conclusion was drawn that subgroups of characteristics that are at higher risk of diabetes are,

PRG, PLASMA, BODY and PEDIGREE.

### 3 Conclusion

In this report, initially, the dataset, 700 total observations, was randomly selected via the help of R. Then, according to the picked dataset, the probability that individual females have diabetes was predicted. The result was

$$\text{REP} = -0.8533 + \text{PRG} \cdot 0.0211 + \text{PLASMA} \cdot 0.00597 - \text{BP} \cdot 0.00213 + \text{BODY} \cdot 0.01297 + \text{PEDIGREE} \cdot 0.1361$$

When the value of REP is close to 1, it means the individual female has higher probability to have diabetes.

When the value of REP is close to 0, it means the individual female has a healthy physical body.

The lower the REP is, the healthier the individual female is. The mean square of error is 2.952.

The next, the higher risk subgroups of characteristics that lead to diabetes were gotten. The unmodified model is PRG, PLASMA, BP, BODY and PEDIGREE.

According to their coefficient, we get that PRG (Number of times pregnant), PLASMA (Plasma glucose concentration in saliva), BODY (Body mass index) and PEDIGREE (Diabetes pedigree function) have positive influence on having diabetes. This means that the larger those variables are, the higher probability the individual female has diabetes. The mean square of error is 3.21068. Hence, the final answer is 4 variables PRG, PLASMA, BODY and PEDIGREE.

Some suggestion to prevent diabetes are diet control and doing exercise. Diet control and doing exercise can help people lower their weight. If weight is reduced, so is Body mass index, finally the REP decreases.

# Appendix 1

**Table 1: Data Exploration**

PRG	PLASMA	BP	THICK
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
NA's :5	NA's :5	NA's :5	NA's :5
INSULIN	BODY	PEDIGREE	AGE
Min. : 0.0	Min. : 0.00	Min. :0.0780	Min. :21.00
1st Qu.: 0.0	1st Qu.:27.30	1st Qu.:0.2437	1st Qu.:24.00
Median : 30.5	Median :32.00	Median :0.3725	Median :29.00
Mean : 79.8	Mean :31.99	Mean :0.4719	Mean :33.24
3rd Qu.:127.2	3rd Qu.:36.60	3rd Qu.:0.6262	3rd Qu.:41.00
Max. :846.0	Max. :67.10	Max. :2.4200	Max. :81.00
NA's :5	NA's :5	NA's :5	NA's :5
REP			
Min. :0.000			
1st Qu.:0.000			
Median :0.000			
Mean :0.349			
3rd Qu.:1.000			
Max. :1.000			
NA's :5			

**Table 2: Data Selection**

```

set.seed(2)
RS=sample(1:768,700)
RS=sort(RS)
RS
A=rep(0, 6300)
DIA=matrix(A, 700, 9, byrow = FALSE)
DIA=data.frame(DIA)
for (i in 1:700) {
  DIA[i,]=PIDD[RS[i],]
}
names(DIA)
PRG=DIA[,1]
PLASMA=DIA[,2]
BP=DIA[,3]
THICK=DIA[,4]
INSULIN=DIA[,5]
BODY=DIA[,6]
PEDIGREE=DIA[,7]
AGE=DIA[,8]

```

```
REP=DIA[,9]
PID=data.frame(PRG, PLASMA, BP, THICK, INSULIN, BODY, PEDIGREE, AGE, REP)
names(PID)
```

```
> names(PID)
[1] "PRG"      "PLASMA"    "BP"        "THICK"     "INSULIN"   "BODY"
[7] "PEDIGREE" "AGE"       "REP"
```

### Table 3: Generalize linear model

```
FGLM=glm(REP~.,family=binomial, data=PID)
summary(FGLM)
```

```
> summary(FGLM)

Call:
glm(formula = REP ~ ., family = binomial, data = PID)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5179  -0.7444  -0.4237   0.7737   2.8893

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.2184518  0.7320929 -11.226  < 2e-16 ***
PRG           0.1262473  0.0332820   3.793 0.000149 ***
PLASMA        0.0347246  0.0038264   9.075  < 2e-16 ***
BP           -0.0122318  0.0053997  -2.265 0.023495 *
THICK         0.0026325  0.0071535   0.368 0.712875
INSULIN      -0.0013447  0.0009214  -1.459 0.144442
BODY          0.0863623  0.0156609   5.515 3.5e-08 ***
PEDIGREE      0.8732767  0.3063937   2.850 0.004369 **
AGE           0.0129725  0.0097037   1.337 0.181269
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 913.63  on 699  degrees of freedom
Residual deviance: 672.22  on 691  degrees of freedom
AIC: 690.22

Number of Fisher Scoring iterations: 5
```

**Table 4: Forward Stepwise Selection**

```
MDSF=step(glm(REP~PRG,family=binomial, data = PID),
           list(upper=~PRG+PLASMA+BP+THICK+INSULIN+BODY+PEDIGREE+AGE),
           direction='forward')
summary(MDSF)
MFTP=predict(MDSF, PID)
mean((REP-MFTP)^2)
```

```
> MDSF=step(glm(REP~PRG,family=binomial, data = PID),
+           list(upper=~PRG+PLASMA+BP+THICK+INSULIN+BODY+PEDIGREE+AGE),
+           direction='forward')
```

```
Start: AIC=884.29
REP ~ PRG
```

	Df	Deviance	AIC
+ PLASMA	1	724.89	730.89
+ BODY	1	812.67	818.67
+ PEDIGREE	1	857.86	863.86
+ INSULIN	1	865.30	871.30
+ AGE	1	866.90	872.90
+ THICK	1	872.04	878.04
<none>		880.29	884.29
+ BP	1	878.55	884.55

```
Step: AIC=730.89
REP ~ PRG + PLASMA
```

	Df	Deviance	AIC
+ BODY	1	689.15	697.15
+ PEDIGREE	1	713.37	721.37
+ THICK	1	721.49	729.49
<none>		724.89	730.89
+ AGE	1	724.27	732.27
+ BP	1	724.58	732.58
+ INSULIN	1	724.68	732.68

```
> summary(MDSF)
```

```
Call:
glm(formula = REP ~ PRG + PLASMA + BODY + PEDIGREE + BP + INSULIN,
    family = binomial, data = PID)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.5452	-0.7491	-0.4277	0.7745	2.9394

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-8.0307994	0.7132655	-11.259	< 2e-16 ***
PRG	0.1481018	0.0289007	5.125	2.98e-07 ***
PLASMA	0.0356265	0.0037056	9.614	< 2e-16 ***
BODY	0.0861092	0.0150004	5.740	9.44e-09 ***
PEDIGREE	0.9016105	0.3050307	2.956	0.00312 **
BP	-0.0107342	0.0052334	-2.051	0.04026 *
INSULIN	-0.0013186	0.0008272	-1.594	0.11091

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 913.63 on 699 degrees of freedom
Residual deviance: 674.06 on 693 degrees of freedom
AIC: 688.06
```

Number of Fisher Scoring iterations: 5

```
Step: AIC=691.04
REP ~ PRG + PLASMA + BODY + PEDIGREE
```

	Df	Deviance	AIC
+ BP	1	676.60	688.60
+ INSULIN	1	678.30	690.30
<none>		681.04	691.04
+ AGE	1	679.81	691.81
+ THICK	1	680.41	692.41

```
Step: AIC=688.6
```

```
REP ~ PRG + PLASMA + BODY + PEDIGREE + BP
```

	Df	Deviance	AIC
+ INSULIN	1	674.06	688.06
+ AGE	1	674.42	688.42
<none>		676.60	688.60
+ THICK	1	676.40	690.40

```
Step: AIC=688.06
```

```
REP ~ PRG + PLASMA + BODY + PEDIGREE + BP + INSULIN
```

	Df	Deviance	AIC
<none>		674.06	688.06
+ AGE	1	672.35	688.35
+ THICK	1	674.00	690.00

```
> MFTP=predict(MDSF, PID)
> mean((REP-MFTP)^2)
[1] 3.21068
```

**Table 5: Backward Stepwise Selection**

```
MDSB=step(glm(REP~PRG+PLASMA+BP+THICK+INSULIN+BODY+PEDIGREE+AGE
              ,family = binomial, data = PID), direction='backward')
summary(MDSB)
MFTP=predict(MDSB, PID)
mean((REP-MFTP)[test3]^2)
```

```
- -
> MDSB=step(glm(REP~PRG+PLASMA+BP+THICK+INSULIN+BODY+PEDIGREE+AGE,
+             family = binomial, data = PID),direction='backward')
Start: AIC=690.22
REP ~ PRG + PLASMA + BP + THICK + INSULIN + BODY + PEDIGREE +
      AGE

      Df Deviance   AIC
- THICK  1   672.35 688.35
- AGE    1   674.00 690.00
<none>   1   672.22 690.22
- INSULIN 1   674.34 690.34
- BP      1   677.40 693.40
- PEDIGREE 1   680.64 696.64
- PRG     1   687.12 703.12
- BODY    1   706.92 722.92
- PLASMA  1   777.85 793.85
```

```
Step: AIC=688.35
REP ~ PRG + PLASMA + BP + INSULIN + BODY + PEDIGREE + AGE
```

	Df	Deviance	AIC
- AGE	1	674.06	688.06
<none>		672.35	688.35
- INSULIN	1	674.42	688.42
- BP	1	677.41	691.41
- PEDIGREE	1	681.00	695.00
- PRG	1	687.27	701.27
- BODY	1	711.84	725.84
- PLASMA	1	779.00	793.00

```
Step: AIC=688.06
REP ~ PRG + PLASMA + BP + INSULIN + BODY + PEDIGREE
```

	Df	Deviance	AIC
<none>		674.06	688.06
- INSULIN	1	676.60	688.60
- BP	1	678.30	690.30
- PEDIGREE	1	683.11	695.11
- PRG	1	701.59	713.59
- BODY	1	712.27	724.27
- PLASMA	1	793.85	805.85

```
> summary(MDSB)
```

```
Call:
glm(formula = REP ~ PRG + PLASMA + BP + INSULIN + BODY + PEDIGREE,
     family = binomial, data = PID)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5452  -0.7491  -0.4277   0.7745   2.9394
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.0307994  0.7132655 -11.259  < 2e-16 ***
PRG          0.1481018  0.0289007   5.125 2.98e-07 ***
PLASMA       0.0356265  0.0037056   9.614  < 2e-16 ***
BP          -0.0107342  0.0052334  -2.051  0.04026 *
INSULIN     -0.0013186  0.0008272  -1.594  0.11091
BODY         0.0861092  0.0150004   5.740 9.44e-09 ***
PEDIGREE     0.9016105  0.3050307   2.956  0.00312 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 913.63 on 699 degrees of freedom
Residual deviance: 674.06 on 693 degrees of freedom
AIC: 688.06
```

```
Number of Fisher Scoring iterations: 5
```

```
> MFTP=predict(MDSB, PID)
```

```
> mean((REP-MFTP)^2)
```

```
[1] 3.21068
```



## Appendix 2

**Table 6: Original Dataset**

PRG, PLASMA, BP, THICK, INSULIN, BODY, PEDIGREE, AGE

6,148,72,35,0,33.6,0.627,50,1  
1,85,66,29,0,26.6,0.351,31,0  
8,183,64,0,0,23.3,0.672,32,1  
1,89,66,23,94,28.1,0.167,21,0  
0,137,40,35,168,43.1,2.288,33,1  
5,116,74,0,0,25.6,0.201,30,0  
3,78,50,32,88,31.0,0.248,26,1  
10,115,0,0,0,35.3,0.134,29,0  
2,197,70,45,543,30.5,0.158,53,1  
8,125,96,0,0,0,0.232,54,1  
4,110,92,0,0,37.6,0.191,30,0  
10,168,74,0,0,38.0,0.537,34,1  
10,139,80,0,0,27.1,1.441,57,0  
1,189,60,23,846,30.1,0.398,59,1  
5,166,72,19,175,25.8,0.587,51,1  
7,100,0,0,0,30.0,0.484,32,1  
0,118,84,47,230,45.8,0.551,31,1  
7,107,74,0,0,29.6,0.254,31,1  
1,103,30,38,83,43.3,0.183,33,0  
1,115,70,30,96,34.6,0.529,32,1  
3,126,88,41,235,39.3,0.704,27,0  
8,99,84,0,0,35.4,0.388,50,0  
.....  
4,136,70,0,0,31.2,1.182,22,1  
1,121,78,39,74,39.0,0.261,28,0  
3,108,62,24,0,26.0,0.223,25,0  
0,181,88,44,510,43.3,0.222,26,1  
8,154,78,32,0,32.4,0.443,45,1  
1,128,88,39,110,36.5,1.057,37,1  
7,137,90,41,0,32.0,0.391,39,0  
0,123,72,0,0,36.3,0.258,52,1  
1,106,76,0,0,37.5,0.197,26,0  
6,190,92,0,0,35.5,0.278,66,1  
2,88,58,26,16,28.4,0.766,22,0  
9,170,74,31,0,44.0,0.403,43,1  
9,89,62,0,0,22.5,0.142,33,0  
10,101,76,48,180,32.9,0.171,63,0  
2,122,70,27,0,36.8,0.340,27,0  
5,121,72,23,112,26.2,0.245,30,0  
1,126,60,0,0,30.1,0.349,47,1  
1,93,70,31,0,30.4,0.315,23,0

**Table 7: Summary**

<b>STATISTICAL CONSULTING PROGRAM</b>
The Study of Pima Indian Diabetes
<b>PROJECT SUMMARY</b>
Oct. 13. 2016
Jian Sun
Feifang Hu
Feifang Hu
<a href="mailto:feifang@gwu.edu">feifang@gwu.edu</a>
Summary
The Study of Pima Indian Diabetes
Summary statistics as requested by client
Statistical analysis of project hypotheses
Written report to be provided to client
Graphical summaries as requested by client
Contract estimate:
<b>Important:</b>
The SCP assumes you accept the contract estimate with the understanding that the final SCP invoice may include additional charges. You will be informed of the need for any increase prior to the SCP performing...