UNDERSTANDING THE CORRELATION BETWEEN HEALTH PREDICTORS AND DIABETES DIAGNOSIS

PREPARED BY: KIERRA MANUEL
STAT 410, FALL 2021

INTRODUCTION AND BACKGROUND

"RACIAL AND ETHNIC MINORITIES, DEFINED AS AMERICAN INDIANS AND ALASKA NATIVES, BLACK OR AFRICAN AMERICANS, HISPANICS OR LATINOS, AND ASIAN AMERICANS, NATIVE HAWAIIANS, AND OTHER PACIFIC ISLANDERS, HAVE A HIGHER PREVALENCE AND GREATER BURDEN OF DIABETES COMPARED TO WHITES, AND SOME MINORITY GROUPS ALSO HAVE HIGHER RATES OF COMPLICATIONS." ACCORDING TO CLINICAL.DIABETESJOURNAL.ORG. I CHOSE THIS TOPIC SINCE I AM OF AFRICAN AMERICAN DESCENT AND HAVE MANY FAMILY MEMBER WHO HAVE EXPERIENCED SEVERE COMPLICATION WITH DIABETES. I WOULD LIKE TO KNOW IF THERE IS A WAY TO PREDICT IF SOMEONE HAS DIABETES AND HOW WE CAN USE THIS INFORMATION TO PREVENT DIABETES DIAGNOSIS.

DATA DESCRIPTION

THIS DATA CAME FROM THE NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES. SINCE IT IS MEDICALLY IMPOSSIBLE FOR SOME DATA RECORD TO BE 0, I CLEANED THE DATA BY REPLACING 0 VALUES WITH THE AVERAGE OF THE PARTICULAR COLUMN. THE PURPOSE OF THIS STUDY IS TO DIAGNOSTICALLY PREDICT WHETHER OR NOT A PATIENT HAS DIABETES BASED ON CERTAIN HEALTH DIAGNOSTIC MEASUREMENTS. ALL THE PATIENTS IN THIS STUDY ARE FEMALES AT LEAST 21 YEARS OF AGE AND OF THE PRIMIA INDIAN HERITAGE. THE SEVERAL MEDICAL PREDICTOR VARIABLES INCLUDE THE NUMBER OF PREGNANCIES, BMI, INSULIN LEVEL, AGE , BLOOD PRESSURE, SKIN THICKNESS, AND DIABETES PEDIGREE FUNCTION. 768 OBSERVATIONS WERE COLLECTED. THE OUTCOME VARIABLE 0 OR 1 INDICATED WHETHER A PERSON HAS DIABETES OR NOT. THE RESPONSE VARIABLE, OUTCOME OF DIABETES DIAGNOSIS, IS BINARY, THAT IS, TRUE OR FALSE DENOTED 0 AND 1 WHICH IS WHY BINARY LOGISTIC, PROBIT, AND COMPLEMENTARY LOG-LOG MODELS WERE USED FOR THIS ANALYSIS.

RESULTS

AFTER COMPARING THE GOODNESS OF FIT TEST FOR THESE THREE REGRESSION MODELS, PROBIT MODEL HAS THE LOWEST AIC, AICC, AND BIC SCORES MEANING IT HAS THE BEST FIT. ACCORDING TO THE SAS AND R OUTPUT, THE P-VALUE FROM THE DEVIANCE TEST INDICATES THE PROBIT MODEL IS A GOOD FIT. THE RESULTS FROM APPLYING THE PROBIT MODEL INDICATE THAT THE NUMBER OF PREGNANCIES, GLUCOSE LEVELS, BMI, AND DIABETES PEDIGREE FUNCTION ARE PREDICTORS SIGNIFICANT AT THE 5% SIGNIFICANCE LEVEL. THE FITTED MODEL IS PHI(OUTCOME) = -5.3539 + 0.0722*PREGNANCIES + 0.0220*PREGNANCIES + 0.0220*GLUCOSE - 0.0053*BLOODPRESSURE + 0.0027*SKINTHICKNESS - 0.0006*INSULIN + 0.0551*BMI + 0.4425*DIABETESPEDIGREEFUNCTION + 0.0084*AGE . THE ESTIMATED REGRESSION COEFFICIENTS ARE INTERPRETED AS FOLLOWS, AS THE NUMBER OF PREGNANCIES INCREASE, THE Z-SCORE OF THE ESTIMATED PROBABILITY OF OUTCOME OF DIABETES INCREASES BY 0.0722. AS BMI INCREASES THE Z-SCORE OF THE ESTIMATED PROBABILITY OF OUTCOME OF DIABETES INCREASES BY 0.0551 . AS GLUCOSE LEVELS INCREASE, THE Z-SCORE OF THE ESTIMATED PROBABILITY OF OUTCOME OF DIABETES INCREASES BY 0.0220 . AS THE DIABETES PEDIGREE FUNCTION SCORE INCREASES , THE Z-SCORE OF THE ESTIMATED PROBABILITY OF OUTCOME OF DIABETES INCREASES BY 0.4425.

CONCLUSION

THERE ARE PREDICTORS THAT ARE SIGNIFICANT IN DETERMINING IF SOMEONE HAS DIABETES. THESE PREDICTORS INCLUDE THE NUMBER OF PREGNANCIES, GLUCOSE LEVELS, BMI, AND DIABETES PEDIGREE FUNCTION. THESE PREDICTORS ARE SIGNIFICANT AT THE 5% SIGNIFICANCE LEVEL. THE BEST MODELS FOR THIS DATA ARE BINARY LOGISTIC, PROBIT, AND COMPLEMENTARY LOG-LOG. OUT OF THESE THREE MODELS, THE PROBIT MODEL HAS THE BEST FIT WITH HAVING THE LOWEST GOODNESS OF FIT SCORES. IF WE CAN MANAGE THESE FOUR PREDICTORS, WE WILL BE ABLE TO BRING DOWN DIABETES DIAGNOSIS.

APPENDIX

```
data diabetes;
input pregnancies glocose bloodpressure skinthickness insulin BMI diabetespedfun age outcome @@;
cards;
6   148 72  35  155.55  33.6    0.627   50  1
1   85  66  29  155.55  26.6    0.351   31  0
8   183 64  29.15   155.55  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  29.15   155.55  25.6    0.201   30  0
3   78  50  32  88  31  0.248   26  1
10  115 72.4    29.15   155.55  35.3    0.134   29  0
2   197 70  45  543 30.5    0.158   53  1
8   125 96  29.15   155.55  32.46   0.232   54  1
4   110 92  29.15   155.55  37.6    0.191   30  0
10  168 74  29.15   155.55  38  0.537   34  1
10  139 80  29.15   155.55  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 72.4    29.15   155.55  30  0.484   32  1
0   118 84  47  230 45.8    0.551   31  1
7   107 74  29.15   155.55  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  29.15   155.55  35.4    0.388   50  0
7   196 90  29.15   155.55  39.8    0.451   41  1
9   119 80  35  155.55  29  0.263   29  1
11  143 94  33  146 36.6    0.254   51  1
10  125 70  26  115 31.1    0.205   41  1
7   147 76  29.15   155.55  39.4    0.257   43  1
1   97  66  15  140 23.2    0.487   22  0
13  145 82  19  110 22.2    0.245   57  0
5   117 92  29.15   155.55  34.1    0.337   38  0
5   109 75  26  155.55  36  0.546   60  0
3   158 76  36  245 31.6    0.851   28  1
3   88  58  11  54  24.8    0.267   22  0

13  126 90  29.15   155.55  43.4    0.583   42  1
4   129 86  20  270 35.1    0.231   23  0
1   79  75  30  155.55  32  0.396   22  0
1   122.68  48  20  155.55  24.7    0.14    22  0
7   62  78  29.15   155.55  32.6    0.391   41  0
5   95  72  33  155.55  37.7    0.37    27  0
0   131 72.4    29.15   155.55  43.2    0.27    26  1
2   112 66  22  155.55  25  0.307   24  0
3   113 44  13  155.55  22.4    0.14    22  0
2   74  72.4    29.15   155.55  32.46   0.102   22  0
7   83  78  26  71  29.3    0.767   36  0
0   101 65  28  155.55  24.6    0.237   22  0
5   137 108 29.15   155.55  48.8    0.227   37  1
2   110 74  29  125 32.4    0.698   27  0
13  106 72  54  155.55  36.6    0.178   45  0
2   100 68  25  71  38.5    0.324   26  0
15  136 70  32  110 37.1    0.153   43  1
1   107 68  19  155.55  26.5    0.165   24  0
1   80  55  29.15   155.55  19.1    0.258   21  0
4   123 80  15  176 32  0.443   34  0
7   81  78  40  48  46.7    0.261   42  0
4   134 72  29.15   155.55  23.8    0.277   60  1
2   142 82  18  64  24.7    0.761   21  0
6   144 72  27  228 33.9    0.255   40  0
2   92  62  28  155.55  31.6    0.13    24  0
1   71  48  18  76  20.4    0.323   22  0
6   93  50  30  64  28.7    0.356   23  0
1   122 90  51  220 49.7    0.325   31  1
1   163 72  29.15   155.55  39  1.222   33  1
1   151 60  29.15   155.55  26.1    0.179   22  0
0   125 96  29.15   155.55  22.5    0.262   21  0
1   81  72  18  40  26.6    0.283   24  0
2   85  65  29.15   155.55  39.6    0.93    27  0
1   126 56  29  152 28.7    0.801   21  0
1   96  122 29.15   155.55  22.4    0.207   27  0
4   144 58  28  140 29.5    0.287   37  0
```

5	109	62	41	129	35.8	0.514	25	1	
6	125	68	30	120	30	0.464	32	0	
5	85	74	22	155.55	29	1.224	32	1	
5	112	66	29.15	155.55	37.8	0.261	41	1	
0	177	60	29	478	34.6	1.072	21	1	
2	158	90	29.15	155.55	31.6	0.805	66	1	
7	119	72.4	29.15	155.55	25.2	0.209	37	0	
7	142	60	33	190	28.8	0.687	61	0	
1	100	66	15	56	23.6	0.666	26	0	
1	87	78	27	32	34.6	0.101	22	0	
0	101	76	29.15	155.55	35.7	0.198	26	0	
3	162	52	38	155.55	37.2	0.652	24	1	
4	197	70	39	744	36.7	2.329	31	0	
0	117	80	31	53	45.2	0.089	24	0	
4	142	86	29.15	155.55	44	0.645	22	1	
6	134	80	37	370	46.2	0.238	46	1	
1	79	80	25	37	25.4	0.583	22	0	
4	122	68	29.15	155.55	35	0.394	29	0	
3	74	68	28	45	29.7	0.293	23	0	
4	171	72	29.15	155.55	43.6	0.479	26	1	
7	181	84	21	192	35.9	0.586	51	1	
0	179	90	27	155.55	44.1	0.686	23	1	
9	164	84	21	155.55	30.8	0.831	32	1	
0	104	76	29.15	155.55	18.4	0.582	27	0	
1	91	64	24	155.55	29.2	0.192	21	0	
4	91	70	32	88	33.1	0.446	22	0	
3	139	54	29.15	155.55	25.6	0.402	22	1	
6	119	50	22	176	27.1	1.318	33	1	
2	146	76	35	194	38.2	0.329	29	0	
9	184	85	15	155.55	30	1.213	49	1	
10	122	68	29.15	155.55	31.2	0.258	41	0	
0	165	90	33	680	52.3	0.427	23	0	
9	124	70	33	402	35.4	0.282	34	0	
1	111	86	19	155.55	30.1	0.143	23	0	
9	106	52	29.15	155.55	31.2	0.38	42	0	
2	129	84	29.15	155.55	28	0.284	27	0	

0	137	70	38	155.55	33.2	0.17	22	0	
0	119	66	27	155.55	38.8	0.259	22	0	
7	136	90	29.15	155.55	29.9	0.21	50	0	
4	114	64	29.15	155.55	28.9	0.126	24	0	
0	137	84	27	155.55	27.3	0.231	59	0	
2	105	80	45	191	33.7	0.711	29	1	
7	114	76	17	110	23.8	0.466	31	0	
8	126	74	38	75	25.9	0.162	39	0	
4	132	86	31	155.55	28	0.419	63	0	
3	158	70	30	328	35.5	0.344	35	1	
0	123	88	37	155.55	35.2	0.197	29	0	
4	85	58	22	49	27.8	0.306	28	0	
0	84	82	31	125	38.2	0.233	23	0	
0	145	72.4	29.15	155.55	44.2	0.63	31	1	
0	135	68	42	250	42.3	0.365	24	1	
1	139	62	41	480	40.7	0.536	21	0	
0	173	78	32	265	46.5	1.159	58	0	
4	99	72	17	155.55	25.6	0.294	28	0	
8	194	80	29.15	155.55	26.1	0.551	67	0	
2	83	65	28	66	36.8	0.629	24	0	
2	89	90	30	155.55	33.5	0.292	42	0	
4	99	68	38	155.55	32.8	0.145	33	0	
4	125	70	18	122	28.9	1.144	45	1	
3	80	72.4	29.15	155.55	32.46	0.174	22	0	
6	166	74	29.15	155.55	26.6	0.304	66	0	
5	110	68	29.15	155.55	26	0.292	30	0	
2	81	72	15	76	30.1	0.547	25	0	
7	195	70	33	145	25.1	0.163	55	1	
6	154	74	32	193	29.3	0.839	39	0	
2	117	90	19	71	25.2	0.313	21	0	
3	84	72	32	155.55	37.2	0.267	28	0	
6	122.68	68	41	155.55	39	0.727	41	1	
7	94	64	25	79	33.3	0.738	41	0	
3	96	78	39	155.55	37.3	0.238	40	0	
10	75	82	29.15	155.55	33.3	0.263	38	0	
0	180	90	26	90	36.5	0.314	35	1	

```
*fit binomial regression model;
proc genmod;
model outcome(event="1") = pregnancies glocose bloodpressure skinthickness insulin BMI diabetespedfun age / dist = binomial link = probit;
run;

*checking model fit;
proc genmod;
model outcome = / dist=binomial link=probit;
run;

data deviance_test;
deviance = -2*(-496.7420 - (-356.6327));
pvalue = 1 - probchi(deviance, 8);
run;

proc print noobs;
run;
```

```
*use fitted model for prediction;
data predict;
input pregnancies glocose bloodpressure skinthickness insulin BMI diabetespedfun age;
cards;
10 168 74 29.15 155.55 38 0.537 34
;
run;

data diabetes;
set diabetes predict;
run;

proc genmod;
model outcome(event="1") = pregnancies glocose bloodpressure skinthickness insulin BMI diabetespedfun age / dist = binomial link = probit;
output out=outdata p=presponse;
run;
```

```
*use fitted model for prediction;
data predict;
input pregnancies glocose bloodpressure skinthickness insulin BMI diabetespedfun age;
cards;
10 168 74 29.15 155.55 38 0.537 34
;
run;

data diabetes;
set diabetes predict;
run;

proc genmod;
model outcome(event="1") = pregnancies glocose bloodpressure skinthickness insulin BMI diabetespedfun age / dist = binomial link = probit;
output out=outdata p=presponse;
run;

proc print data=outdata (firstobs=769) noobs;
var presponse;
run;
```

The SAS System

The GENMOD Procedure

Model Information	
Data Set	WORK.DIABETES
Distribution	Binomial
Link Function	Probit
Dependent Variable	outcome

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-355.9524	
Full Log Likelihood		-355.9524	
AIC (smaller is better)		729.9048	
AICC (smaller is better)		730.1423	
BIC (smaller is better)		771.6990	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.3539	0.4484	-6.2328	-4.4750	142.55	<.0001
pregnancies	1	0.0722	0.0183	0.0362	0.1081	15.49	<.0001
glocose	1	0.0220	0.0022	0.0177	0.0263	102.26	<.0001
bloodpressure	1	-0.0053	0.0050	-0.0152	0.0045	1.13	0.2882
skinthickness	1	0.0027	0.0077	-0.0124	0.0177	0.12	0.7301
insulin	1	-0.0006	0.0007	-0.0019	0.0007	0.75	0.3853
BMI	1	0.0551	0.0103	0.0349	0.0753	28.60	<.0001
diabetespedfun	1	0.4425	0.1626	0.1239	0.7611	7.41	0.0065
age	1	0.0084	0.0055	-0.0024	0.0192	2.33	0.1266
Scale	0	1.0000	0.0000	1.0000	1.0000		

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-496.7420	
Full Log Likelihood		-496.7420	
AIC (smaller is better)		995.4839	
AICC (smaller is better)		995.4891	
BIC (smaller is better)		1000.1277	

The SAS System

deviance	pvalue
280.219	0

The SAS System

presponse
0.89876

##STAT 410 Project

```
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ dplyr   1.0.7
## ✓ tibble  3.1.5     ✓ stringr 1.4.0
## ✓ tidyr   1.1.4     ✓ forcats 0.5.1
## ✓ purrr   0.3.4
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
## Warning: package 'tibble' was built under R version 4.1.1
```

```
## Warning: package 'tidyr' was built under R version 4.1.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Importing diabetes data

```
diabetesexcel <- read_csv("~/Desktop/diabetesexcel.csv")
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   Pregnancies = col_double(),
##   Glucose = col_double(),
##   BloodPressure = col_double(),
##   SkinThickness = col_double(),
##   Insulin = col_double(),
##   BMI = col_double(),
##   DiabetesPedigreeFunction = col_double(),
##   Age = col_double(),
##   Outcome = col_double()
## )
```

```
glimpse(diabetesexcel)
```

```
## Rows: 768
## Columns: 9
## $ Pregnancies              <dbl> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, …
## $ Glucose                  <dbl> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125…
## $ BloodPressure            <dbl> 72.0, 66.0, 64.0, 66.0, 40.0, 74.0, 50.0, 72.…
## $ SkinThickness            <dbl> 35.00, 29.00, 29.15, 23.00, 35.00, 29.15, 32.…
## $ Insulin                  <dbl> 155.55, 155.55, 155.55, 94.00, 168.00, 155.55…
## $ BMI                      <dbl> 33.60, 26.60, 23.30, 28.10, 43.10, 25.60, 31.…
## $ DiabetesPedigreeFunction <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.2…
## $ Age                      <dbl> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 3…
## $ Outcome                  <dbl> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, …
```

Apply Probit Model

```
summary(fitted.model <- glm(Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + BMI + Dia
betesPedigreeFunction + Age, data = diabetesexcel, family = binomial(link=probit)))
```

```
## Call:
## glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure +
##     SkinThickness + Insulin + BMI + DiabetesPedigreeFunction +
##     Age, family = binomial(link = probit), data = diabetesexcel)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6538  -0.7315  -0.3762   0.7338   2.4245
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -5.3538932  0.4467068 -11.985  < 2e-16 ***
## Pregnancies               0.0721851  0.0187424   3.851 0.000117 ***
## Glucose                   0.0219995  0.0021727  10.125  < 2e-16 ***
## BloodPressure            -0.0053349  0.0049651  -1.074 0.282605
## SkinThickness             0.0026553  0.0075884   0.350 0.726398
## Insulin                  -0.0005706  0.0006836  -0.835 0.403851
## BMI                       0.0550963  0.0102032   5.400 6.67e-08 ***
## DiabetesPedigreeFunction  0.4425297  0.1697900   2.606 0.009152 **
## Age                       0.0083976  0.0055885   1.503 0.132929
## ---
## 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: 711.90  on 759  degrees of freedom
## AIC: 729.9
##
## Number of Fisher Scoring iterations: 5
```

Computing AICC for Probit Model

```
p <- 8
n <- 768
print(AICC <- -2*logLik(fitted.model) + 2*p*n/(n-p-1))
```

```
## 'log Lik.' 728.0946 (df=9)
```

Output #BIC

```
BIC(fitted.model)
```

```
## [1] 771.699
```

Checking model fit

```
null.model <- glm(Outcome ~ 1, data = diabetesexcel,
                  family=binomial(link=probit))
print(deviance <- -2*(logLik(null.model) - logLik(fitted.model)))
```

```
## 'log Lik.' 281.5791 (df=1)
```

```
print(p.value <- pchisq(deviance, 8, lower.tail = FALSE))
```

```
## 'log Lik.' 3.409803e-56 (df=1)
```

#using fitted model for prediction

```
print(predict(fitted.model, data.frame(Pregnancies=10, Glucose=168, BloodPressure=74, SkinThickness=29.15, Insuli
n=155.55, BMI=38, DiabetesPedigreeFunction=0.537, Age=34), type = "response"))
```

```
##         1
## 0.8987641
```