Understanding the correlation of health predictors and diabetes diagnosis

By: Kierra Manuel

Today's Agenda

Background	02
Data and Purpose	05
Programming in SAS	07
Programming in R	08
Analysis	10
Interpretations	11
Conclusion	13

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 I have many family member who have experienced 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

Background cont'd

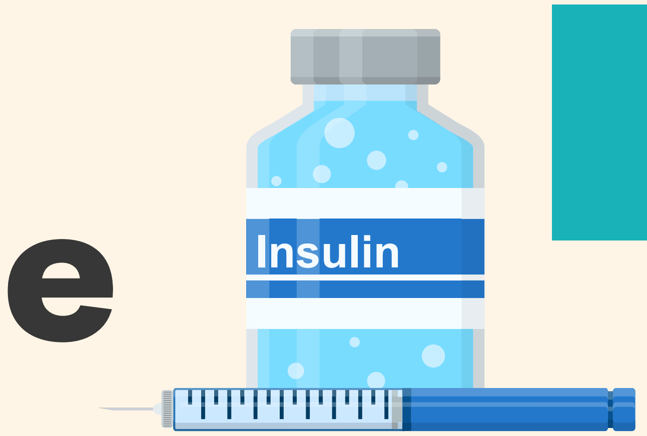
4.9 million African-American adults, or 18.7% of all African Americans ≥ 20 years of age, have diagnosed or undiagnosed diabetes, compared to 7.1% of non-Hispanic white Americans.

The risk of diabetes is 77% higher among African Americans than among non-Hispanic white Americans.

In 2006, African Americans with diabetes were 1.5 times more likely to be hospitalized and 2.3 times more likely to die from diabetes than non-Hispanic whites.

Diabetes Statistics from clinical.diabetesjournal.org

Data and Purpose



This data came from the National Institute of Diabetes and Digestive and Kidney Diseases.

The purpose of this study is to diagnostically pedict whether or not a patient has diabetes based on certain health diagnostic measurements.

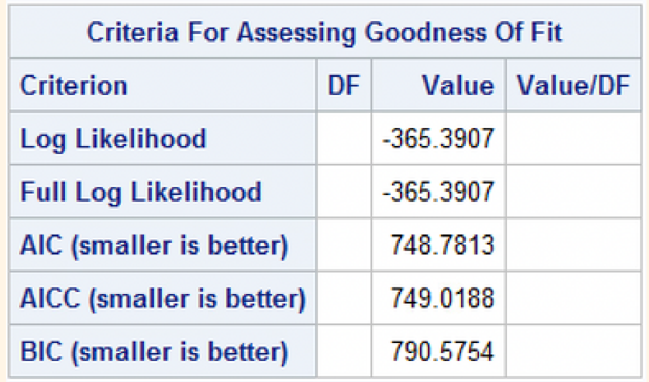
All the patients in this study are females at least 21 years of age and of the Primia Indian Heritage. The several medical predictor variables include the number of pregnancies, BMI, Insulin level, age , blood pressure, skin thickness, and diabetes pedigree function. 768 observations were collected.

The outcome variable 0 or 1 indicated whether a person has diabetes or not.

Data and Purpose cont'd

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.

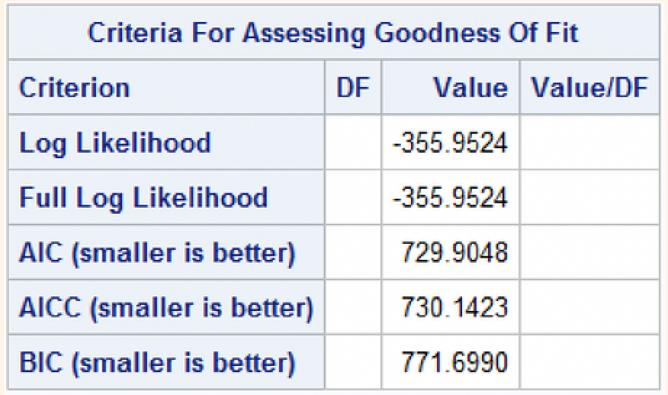
After comparing The Goodness of Fit Test for these three regression models, Probit Model has the lowest AIC, AICC, and BIC scores meaning it has the best fit.



Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Log Likelihood		-365.3907	
Full Log Likelihood		-365.3907	
AIC (smaller is better)		748.7813	
AICC (smaller is better)		749.0188	
BIC (smaller is better)		790.5754	

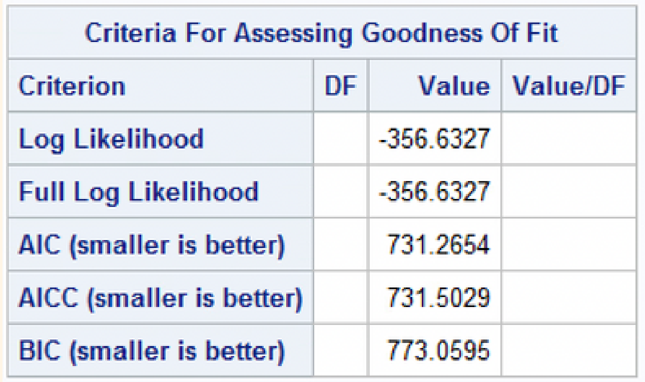
Complementary Log-Log Model



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	

Probit Model

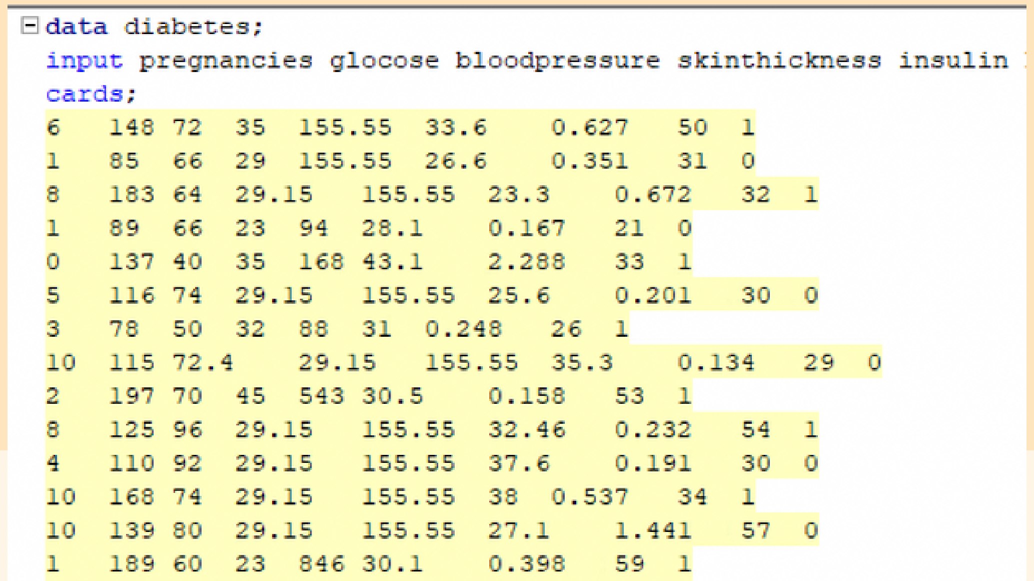


Criteria For Assessing Goodness Of Fit

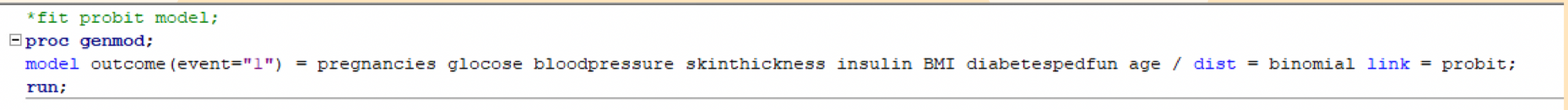
Criterion	DF	Value	Value/DF
Log Likelihood		-356.6327	
Full Log Likelihood		-356.6327	
AIC (smaller is better)		731.2654	
AICC (smaller is better)		731.5029	
BIC (smaller is better)		773.0595	

Binary Logistic Model

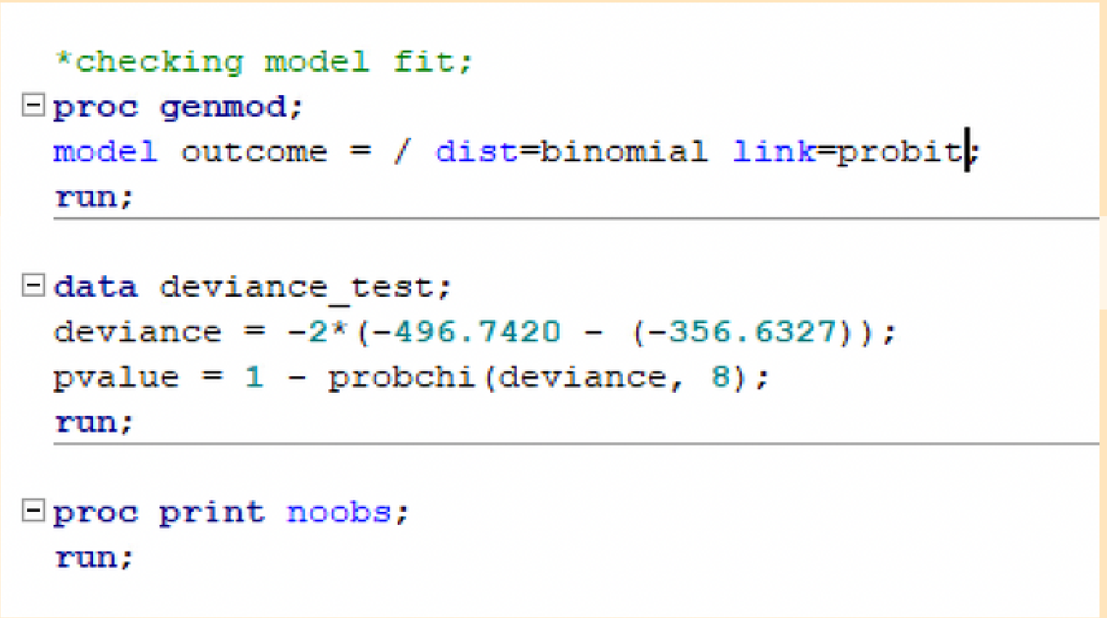
Programming Code in SAS



```
data diabetes;
  input pregnancies glocose bloodpressure skinthickness insulin
  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
```

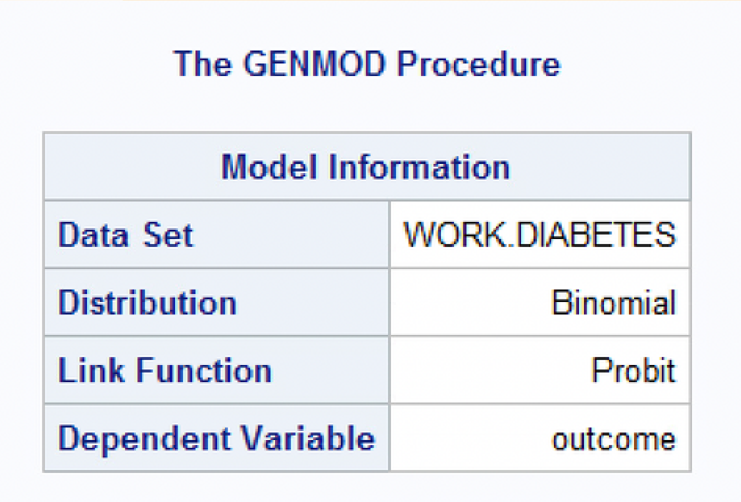


```
 *fit probit model;
proc genmod;
 model outcome(event="1") = pregnancies glocose bloodpressure skinthickness insulin BMI diabetespedfun age / dist = binomial link = probit;
 run;
```



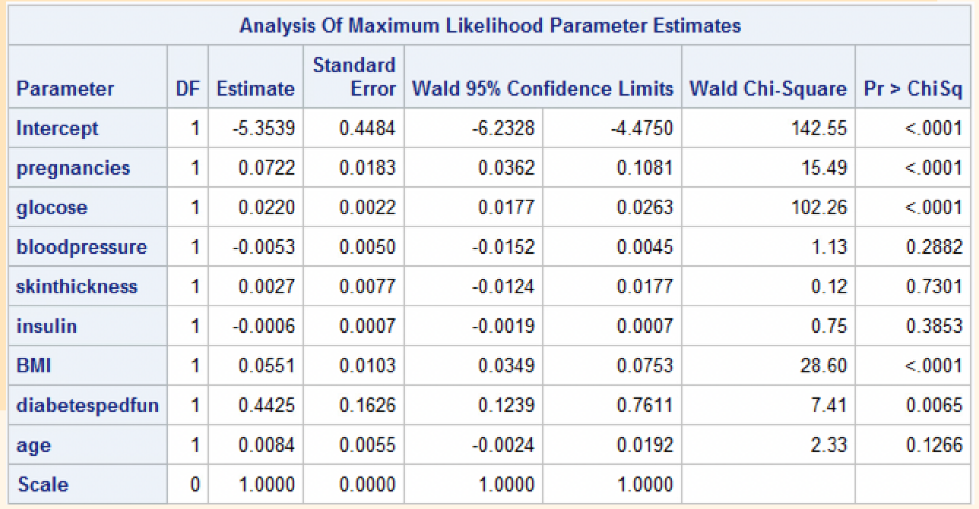
```
 *checking model fit;
proc genmod;
 model outcome = / dist=binomial link=probit;
 run;

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

proc print noobs;
 run;
```

Programming Output in SAS



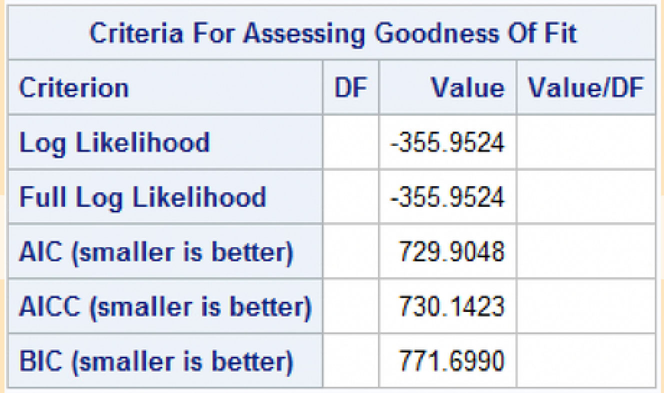
The GENMOD Procedure

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



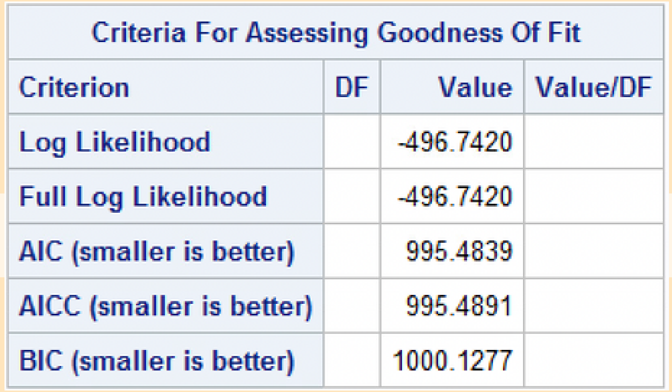
Analysis Of Maximum Likelihood Parameter Estimates

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		



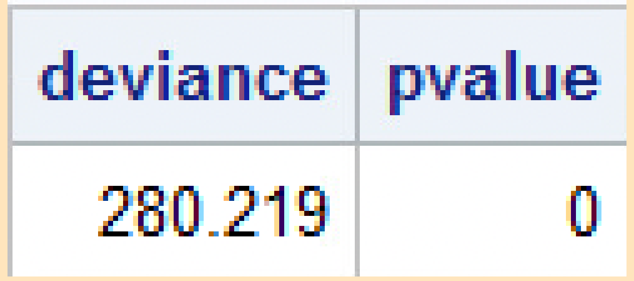
Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Log Likelihood		-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	



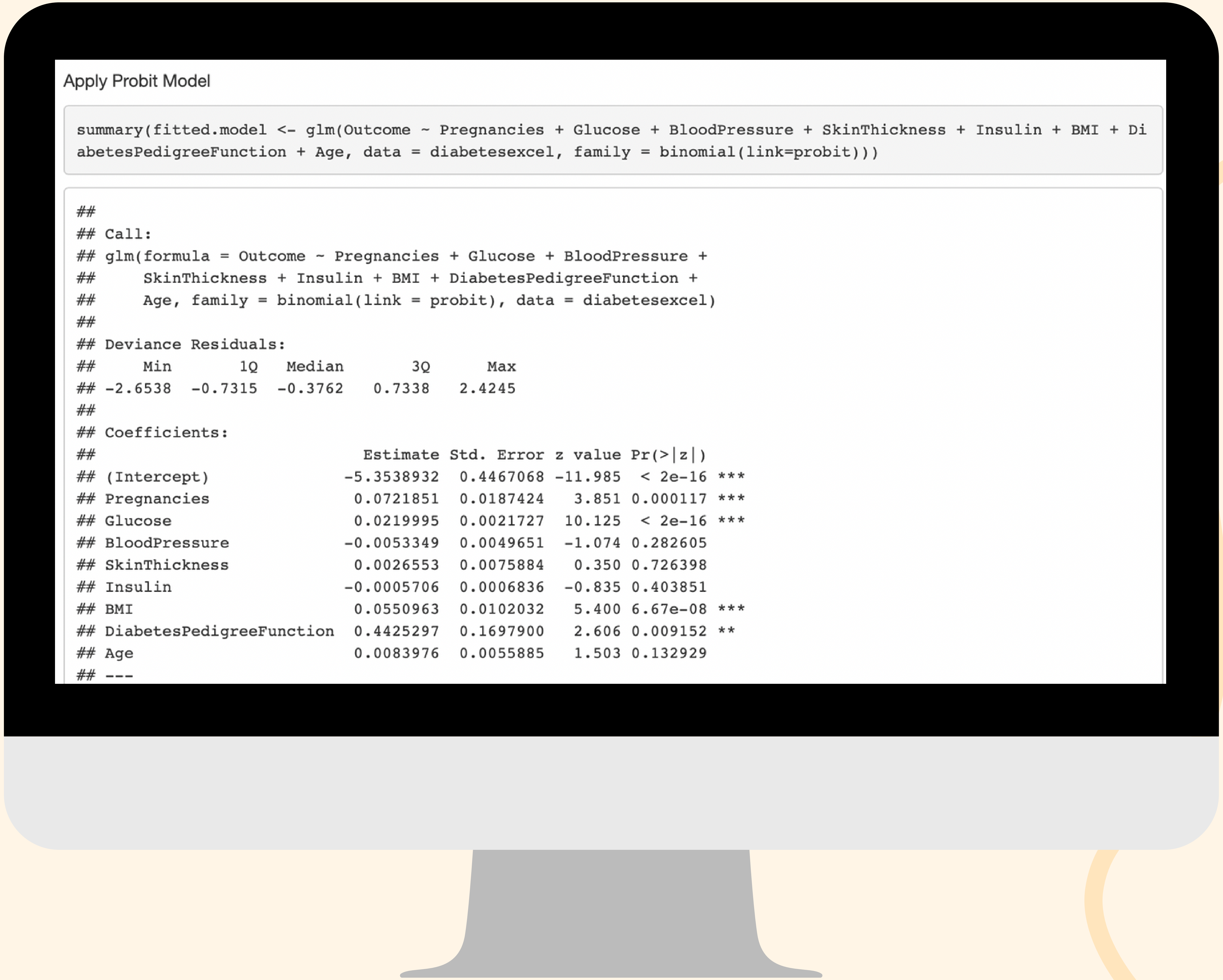
Criteria For Assessing Goodness Of Fit

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	



deviance	pvalue
280.219	0

Programming Code and output in R

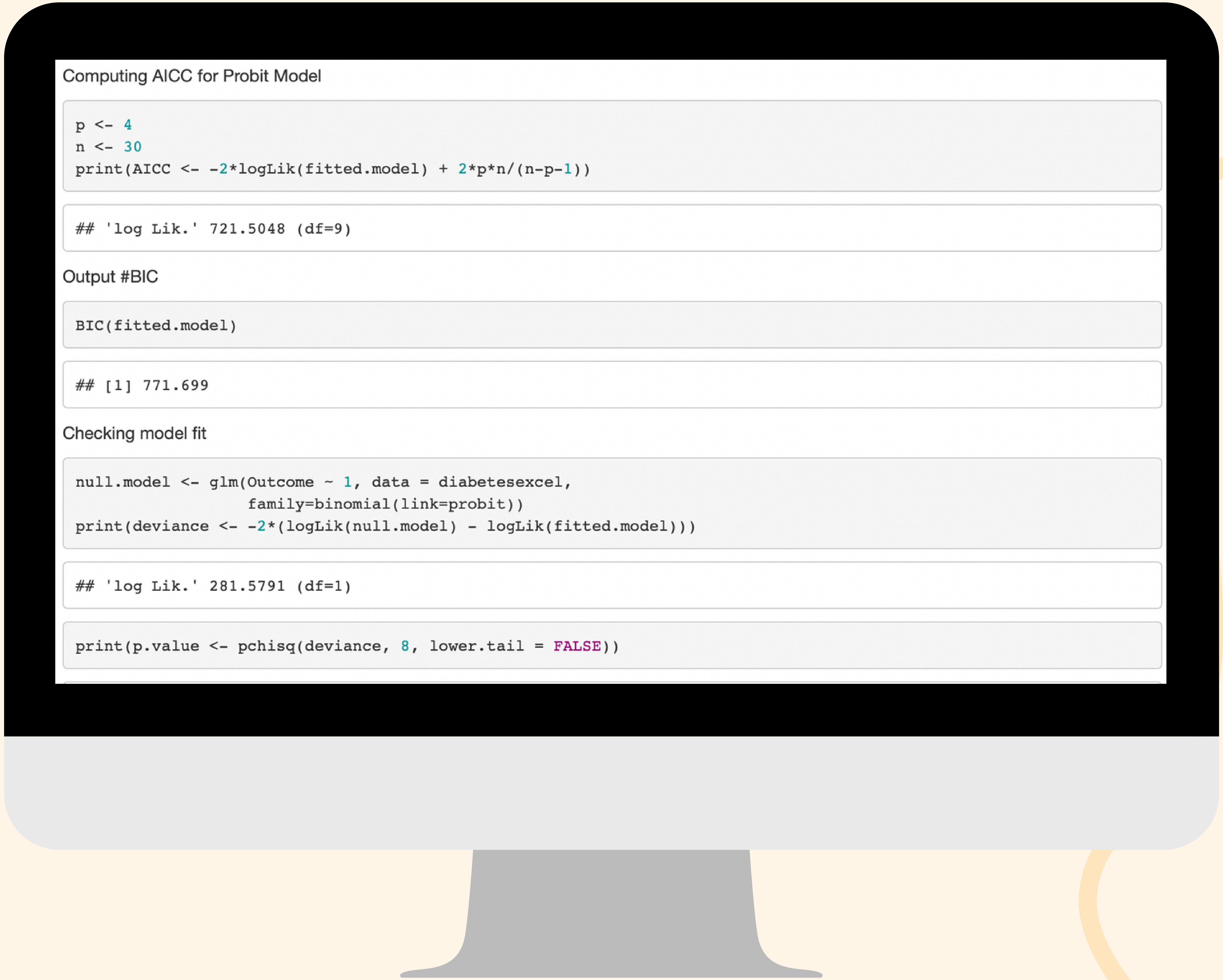


Apply Probit Model

```
summary(fitted.model <- glm(Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + BMI + Di
abetesPedigreeFunction + 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    
## ---
```

Programming Code and output in R cont'd



Computing AICC for Probit Model

```
p <- 4
n <- 30
print(AICC <- -2*logLik(fitted.model) + 2*p*n/(n-p-1))
```

```
## 'log Lik.' 721.5048 (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))
```

Analysis

- According to the SAS and R Output, the p-value from the deviance test indicates the 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*Glucose - 0.0053*BloodPressure + 0.0027*SkinThickness - 0.0006*Insulin + 0.0551*BMI + 0.4425*DiabetesPedigreeFunction + 0.0084*Age

Interpretations

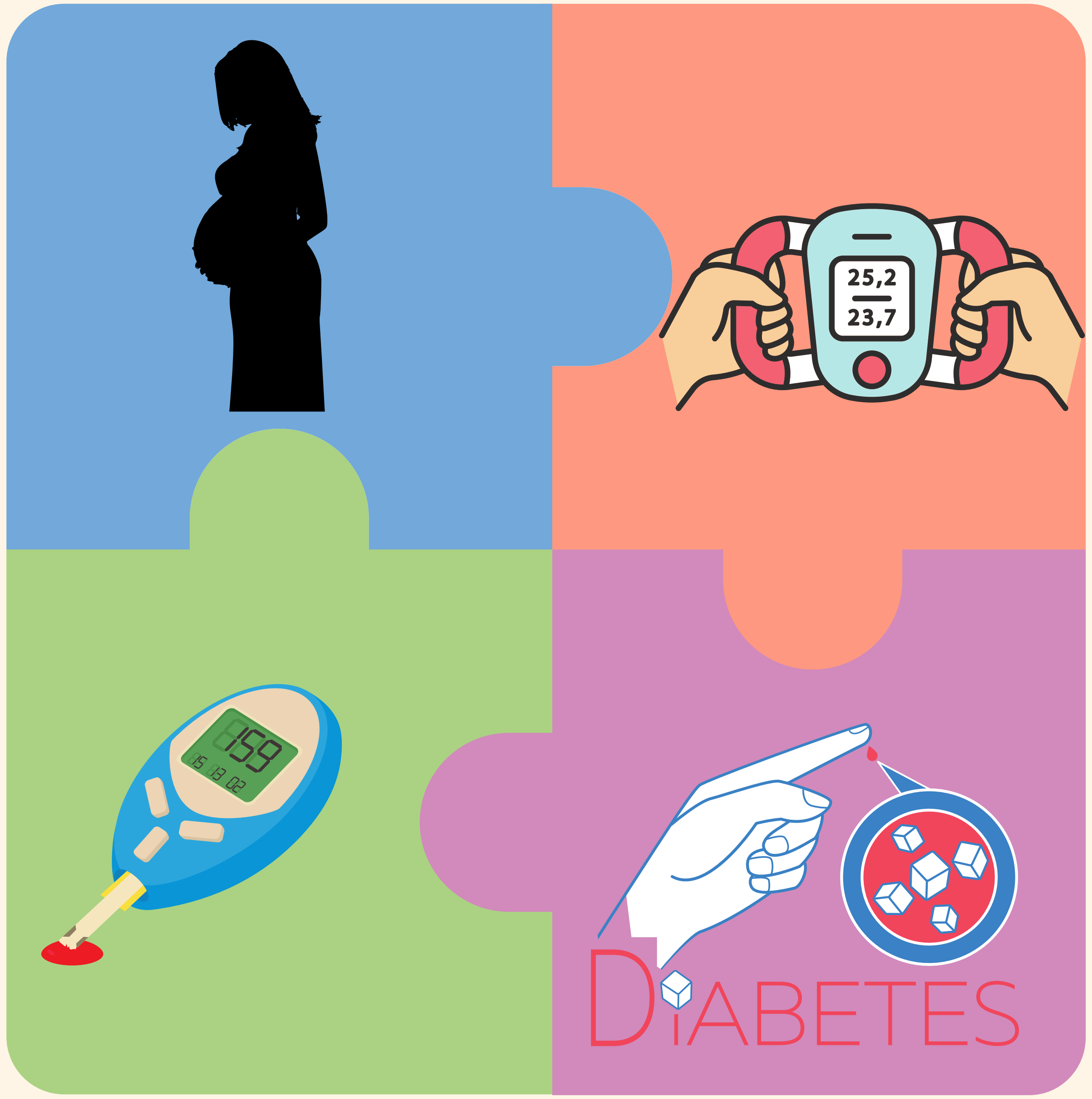
The estimated regression coefficients are interpreted as follows

Number of Pregnancies

As the number of pregnancies increase, the z-score of the estimated probability of outcome of diabetes increases by 0.0722

Glucose Levels

As glucose levels increase, the z-score of the estimated probability of outcome of diabetes increases by 0.0220



BMI

As BMI increases the z-score of the estimated probability of outcome of diabetes increases by 0.0551

Diabetes Pedigree Function

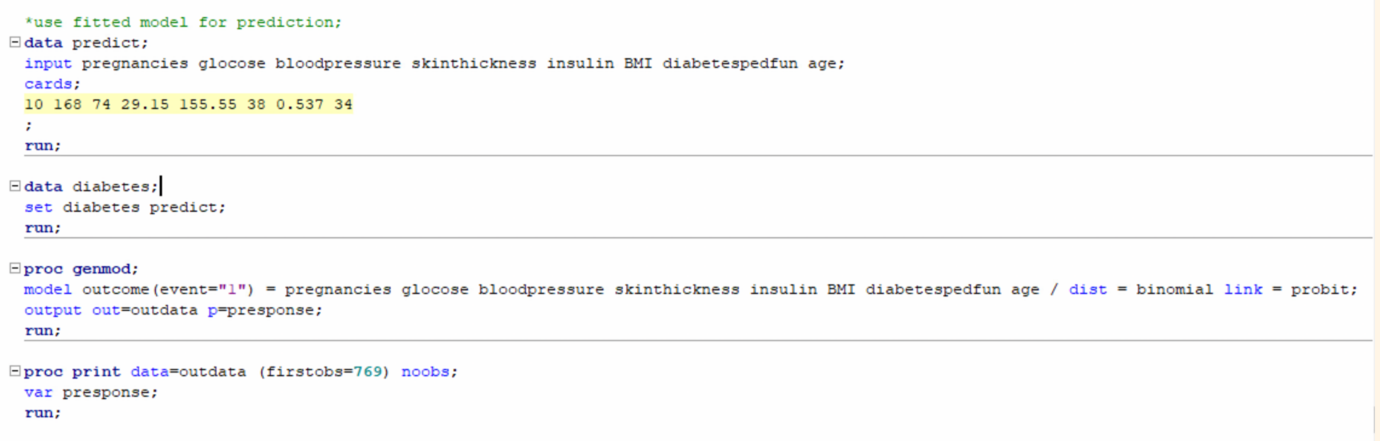
As the diabetes pedigree function score increases , the z-score of the estimated probability of outcome of diabetes increases by 0.4425

Here's a Prediction using health data from a participant in the study that was diagnosed with diabetes.

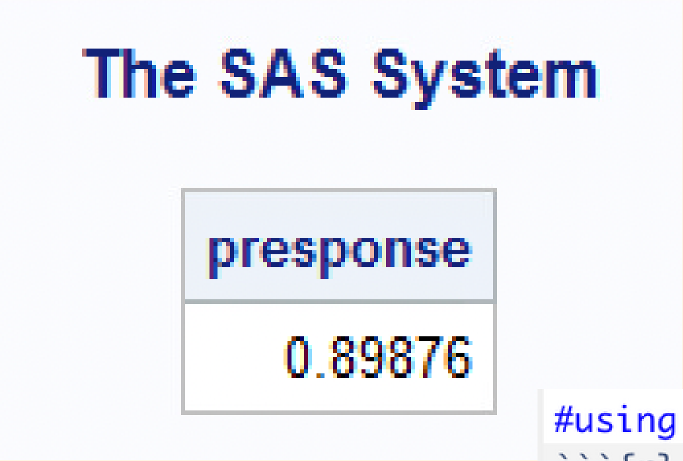
Using fitted model:

P^o = -5.3539 + 0.0722*10 + 0.0220*168 - 0.0053*74 + 0.0027*29.15 - 0.0006*155.55 + 0.0551*38 + 0.4425*0.537 + 0.0084*34 = 1.274298

In SAS:



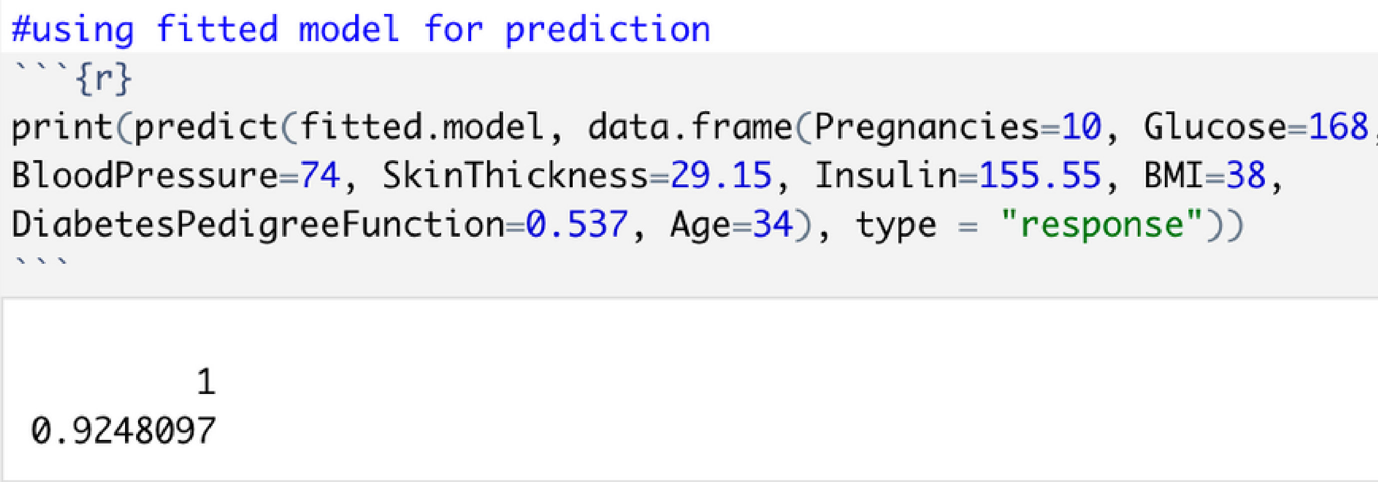
```
*use fitted model for prediction;
data predict;
input pregnancies glocose bloodpressure skinthickness insulin BMI diabetespedfun age;
cards;
10 168 74 29.15 155.55 38 0.537 34
;
run;

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

presponse
0.89876

In R:



~~~
#using fitted model for prediction
```{r}
print(predict(fitted.model, data.frame(Pregnancies=10, Glucose=168
BloodPressure=74, SkinThickness=29.15, Insulin=155.55, BMI=38,
DiabetesPedigreeFunction=0.537, Age=34), type = "response"))
```

        1
0.9248097
~~~

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.

THANK YOU DR. OLGA AND STAT 410 CLASSMATES!