

Diabetes in the Pima Indians

PREDICTING DIABETES ONSET

Allen Baumgarten | STAT5120 | 5/3/2018

# Overview

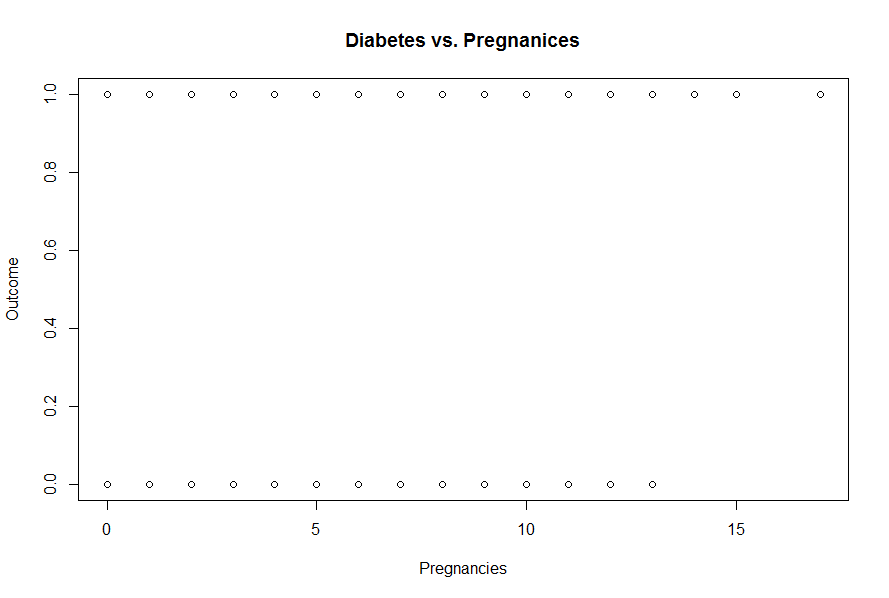
The Pima Indian people are a surviving tribe of native North Americans living in Arizona.[[1]](#footnote-1) Their heritage dates back to at least the late 17th Century if not earlier. Spanish missionaries speak of their tribes and to having made contact with them.[[2]](#footnote-2) Today, the modern descendants of this tribe live in Central Arizona on federally protected land known as the Gila River Indian Community.

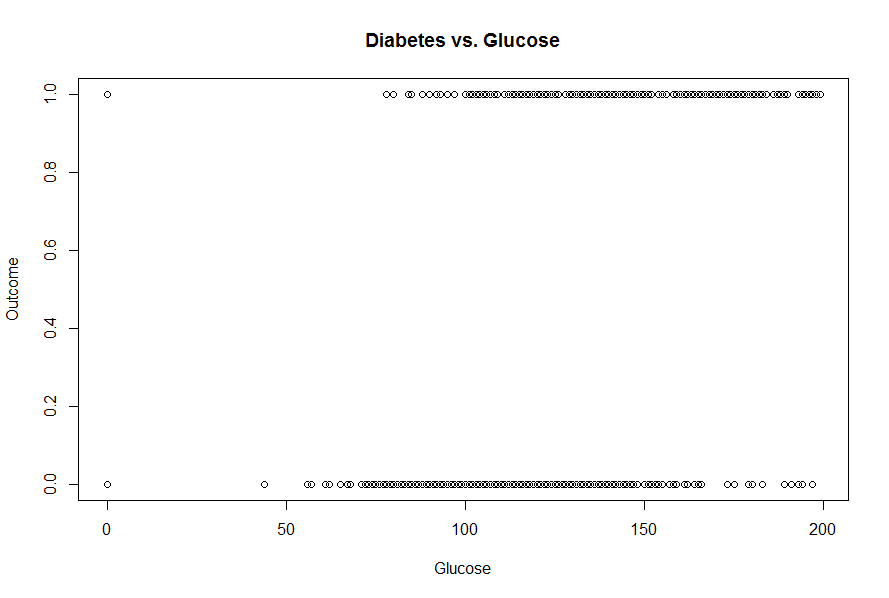
In a fascinating 2004 study, diabetes researchers decided to study this people group as a means of learning more about diabetes onset in a population with minimal genetic variability.

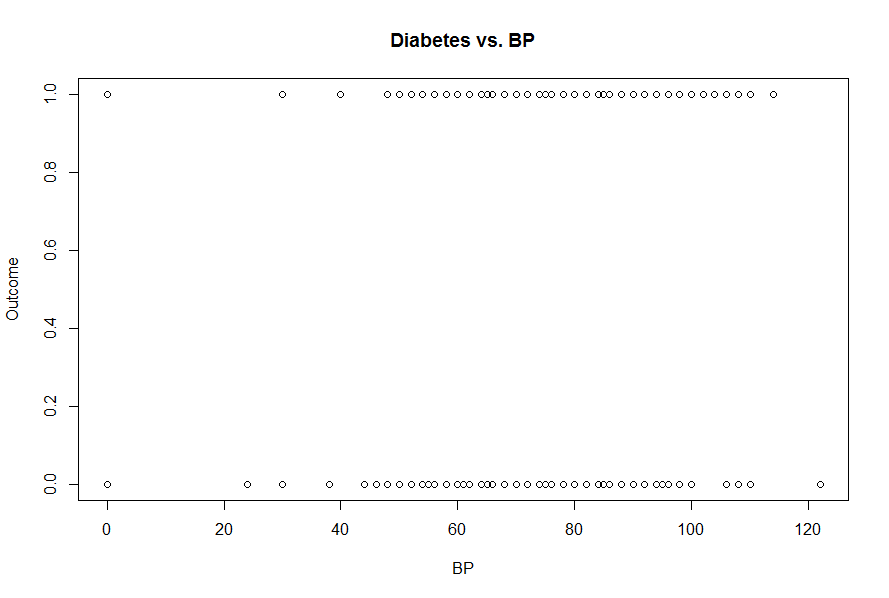
The goal of this study is to ascertain which, if any, predictor variables recorded may provide insights into the likelihood of the pima people developing the onset of diabetes.

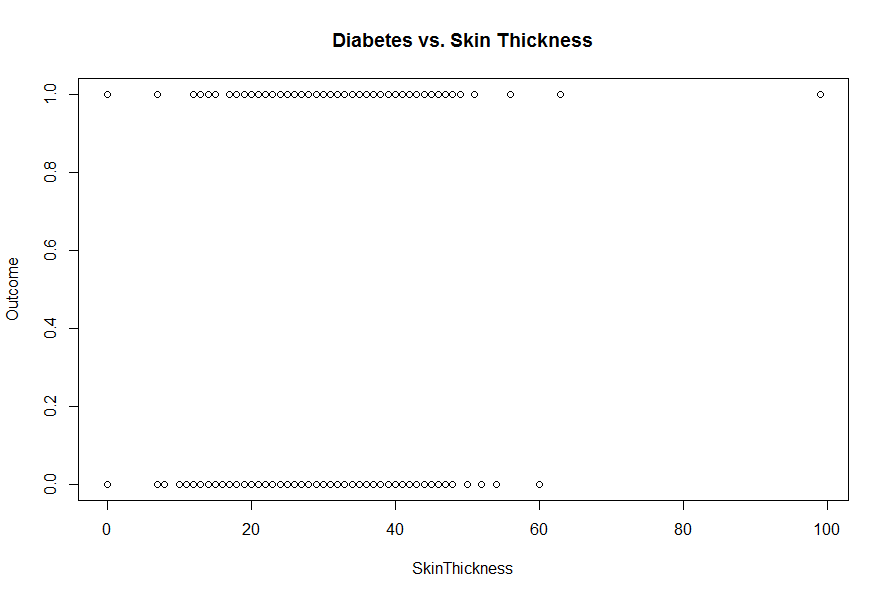
## Variable selection

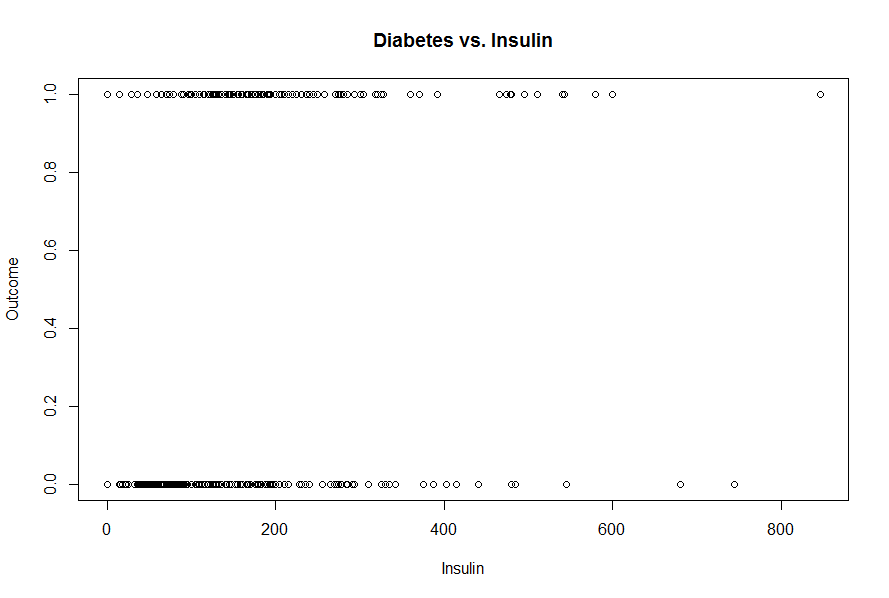
A cursory examination of the predictor variables vs. the outcome of diabetes suggested that Glucose, Insulin, BMI and DiabPedFct are all associated with the outcome of diabetes onset vs. no onset while the other variables were not. Scatter plots of the outcome vs. each of the predictor variables showed that the outcome of diabetes tended to have slightly differing clusters depending on whether these particular measures increased or decreased.

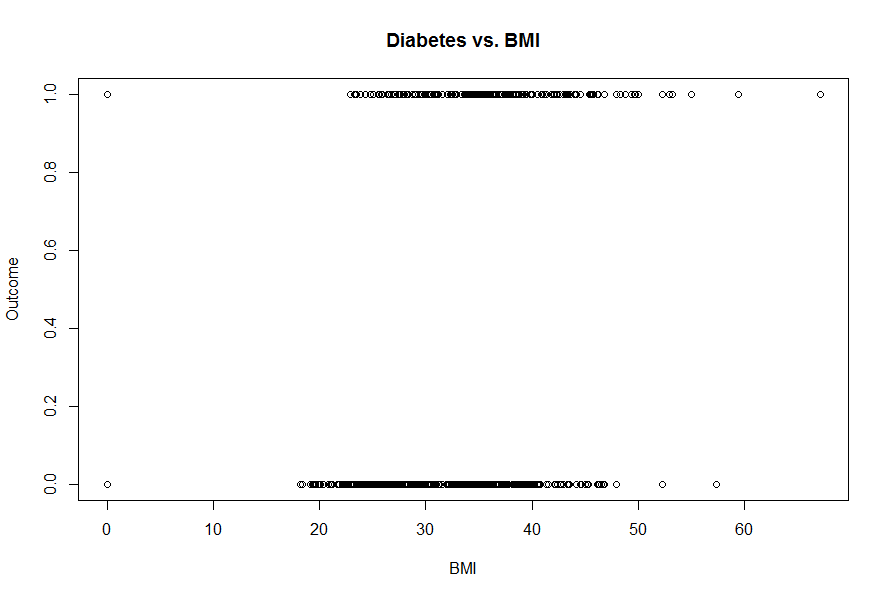
[[3]](#footnote-3)

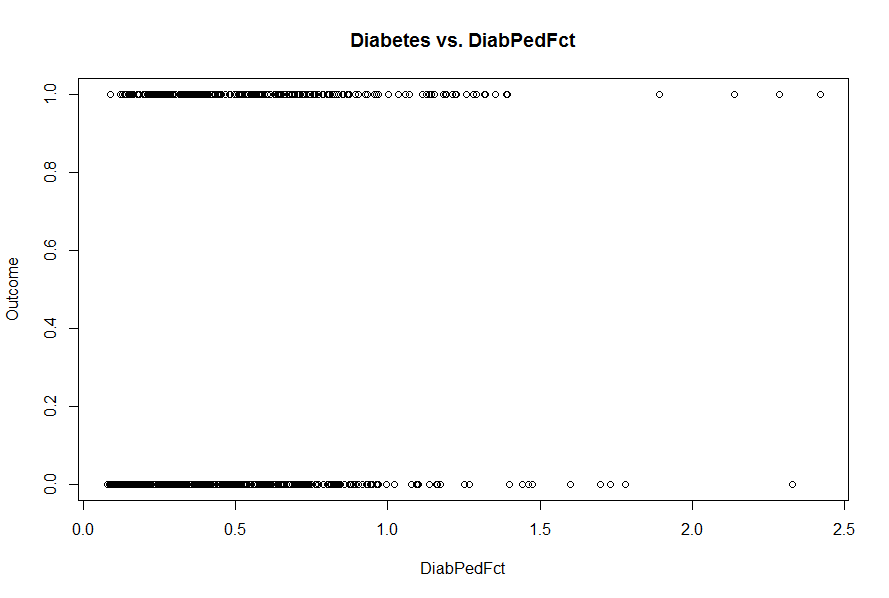


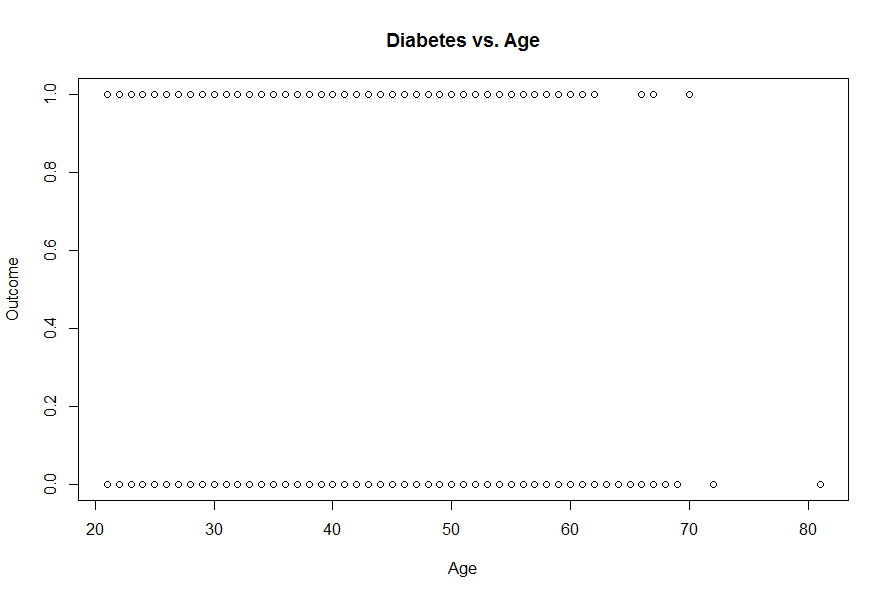






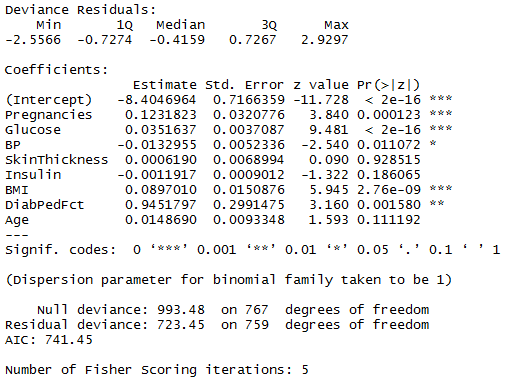




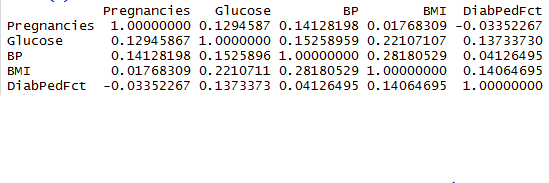


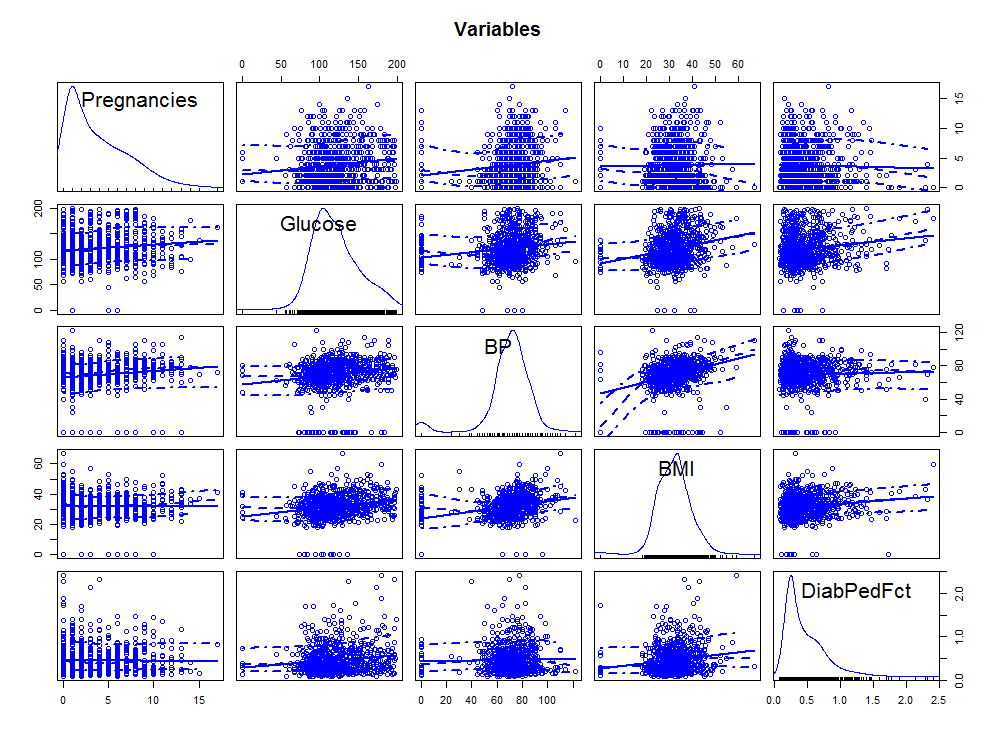
Model Selection Process

A cursory model was developed to investigate which variables might be helpful in predicting diabetes. It was shown that Pregnancies, Glucose, BP, BMI, and DiabPedFct might prove helpful:



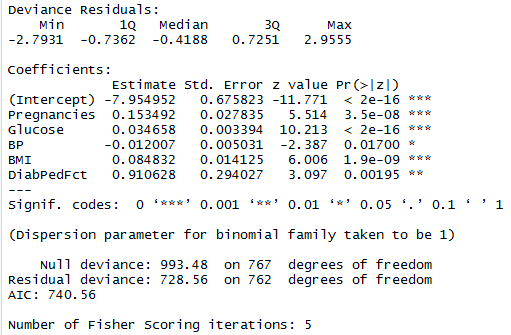
A scatter plot matrix and correlation matrix showed that inclusion of these variables might helpful by having a lack of any strong correlation between them:





Final Model

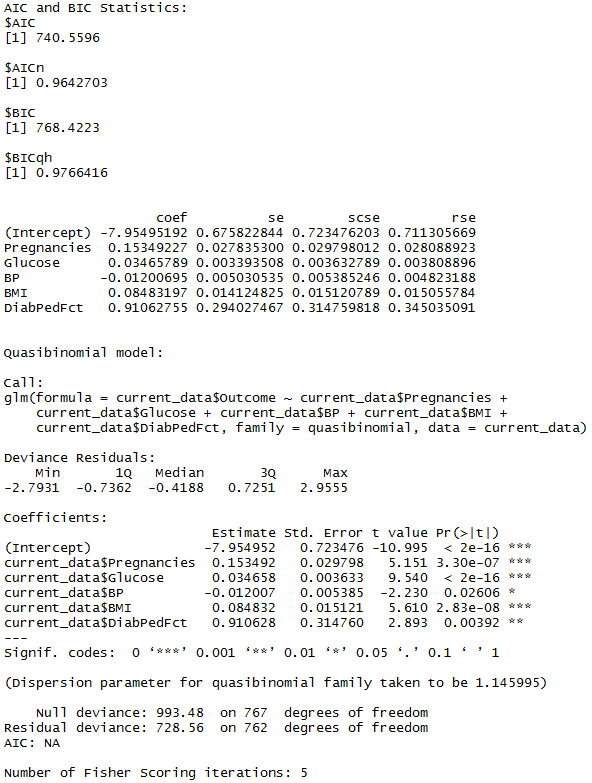
Trimming down to a model with just these predictors was done to investigate whether such a model might be stronger in predicting diabetes onset. Our p-values all indicated statistically strong association with the Outcome variable.

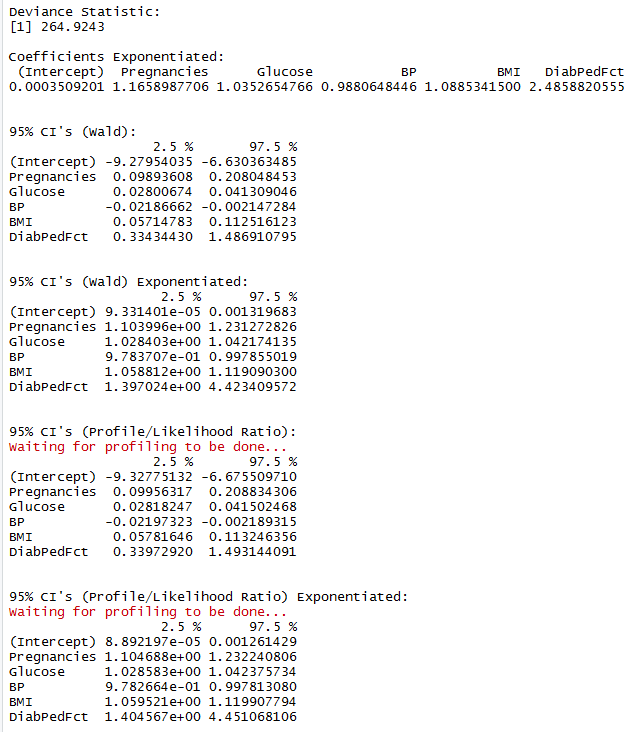


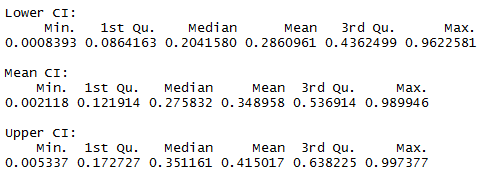
Model Diagnostics

Diagnostic measures were calculated to see if this new reduced model was statistically sound in predicting diabetes outcomes and these diagnostics appear strong.

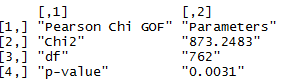
One particular diagnostic is the AIC. The AIC for our reduced model = 740.56 which was slightly better than a fully-loaded model with all variables with an AIC = 741.45.



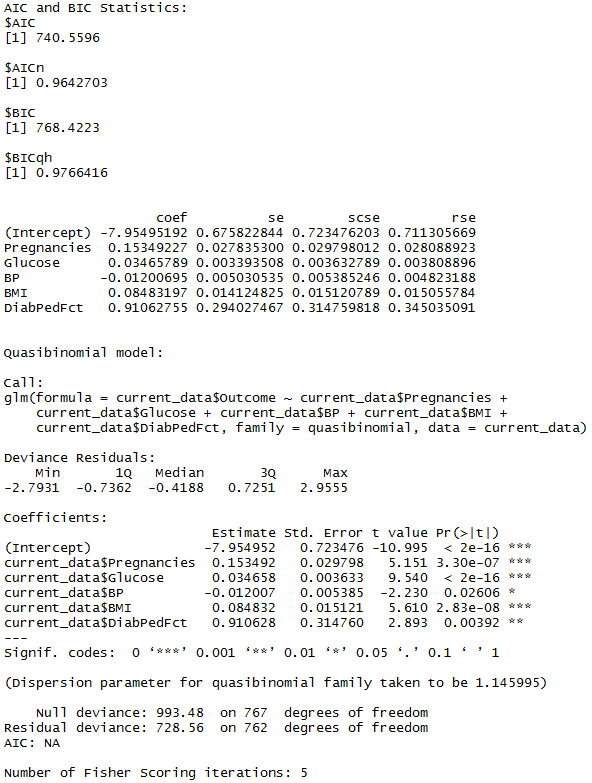


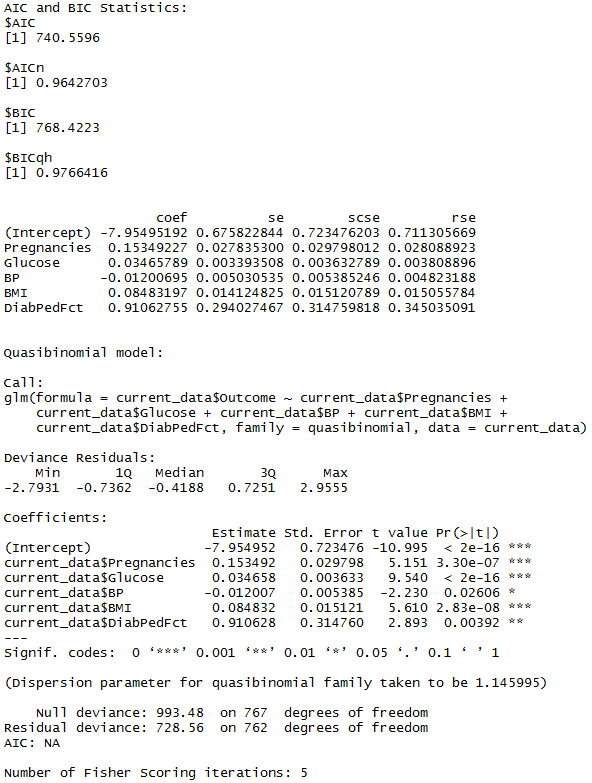


Another diagnostic is the Pearson chi-square goodness-of-fit test. This test divides the Pearson chi-square statistic by the residual degrees of freedom which, if the model fits well, should be close to 1.0. In this case, we find that Chi2/df = 873.25/762 = 1.14, indicating a good fit.



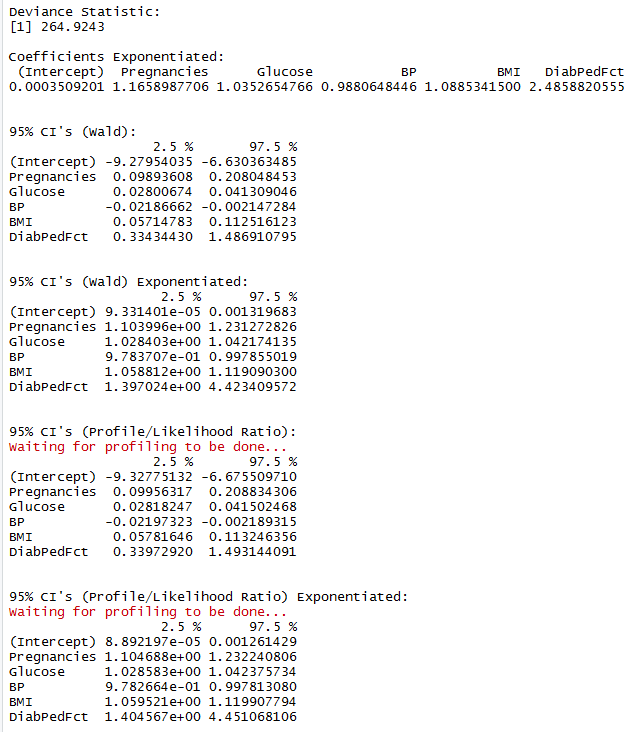
The practice of comparing the standard errors generated by the glm model to scaled and robust standard errors is common. If excess correlation exists, or if there is an underlying probability distribution other than that which we are modeling, our standard errors would be biased. Scaled standard errors (shown in the “scse” column below) are the product of the model standard errors times the square root of the Pearson dispersion. These errors check for correlation between the predictor variables in the model. Here, we see that the scse’s are not much different from our model standard errors. Additionally, the “rse” column shows robust standard errors which, if there are no problems with the model, reduce to the model standard errors. In this case, only intercept errors seem somewhat different; the predictor standard errors are close. Finally, the quasibinomial function is an approach used to reproduce our model in question, albeit with scaled standard errors. This function is shown further below.



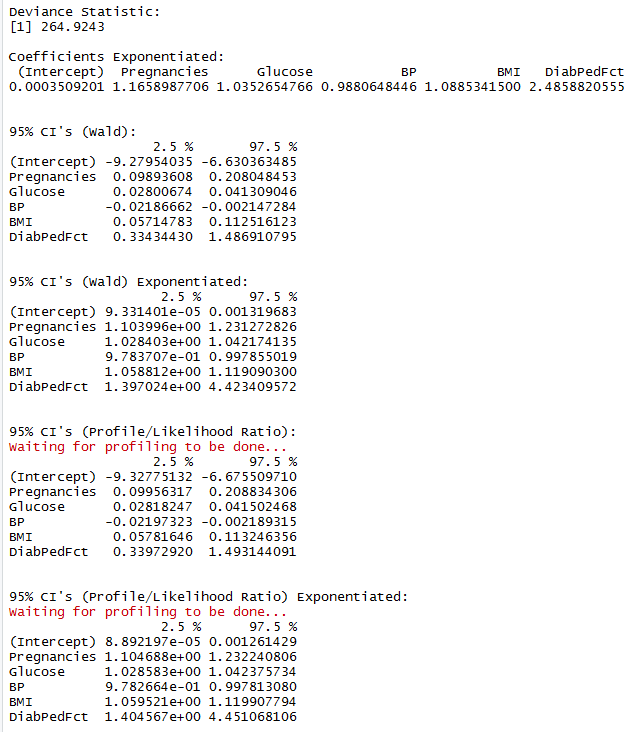


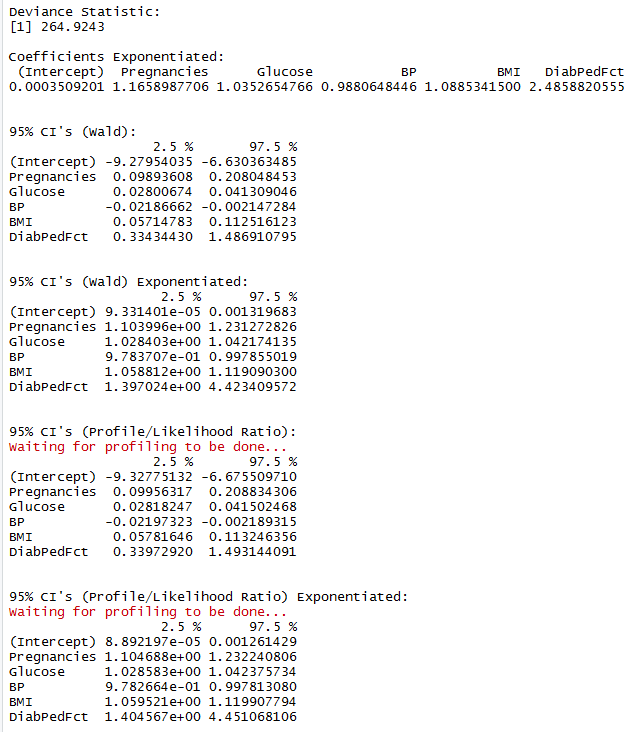
Model Implementation and Interpretation

Our selected model seems sound and we may use it to interpret the odds of diabetes onset. We take our model coefficients shown earlier and exponentiate them:



The exponentiated coefficients with Wald 95% confidence intervals and Profile/Likelihood Ratio 95% confidence intervals are available to provide more accurate interpretation:





Our formal model will then suggest that we can predict diabetes onset in the Pima people group with the following (exponentiated) variables. These variables are shown as odds ratios and provide grounds for interpretation:

Diabetes Onset (yes) = .000 + 1.166(Pregnancy) + 1.035(Glucose) + 0.988(BP) + 1.089(BMI) + 2.486(DiabPedFct)

We can use this model to understand the relationships between our five selected epidemiological measures to understand the likelihood of diabetes onset in this people group. For example, holding all other variables constant, we can be 95% confident that a one-unit change in Pregnancy (yes/no) will result in a 16.6% greater chance (with a range of 10.4% to 23.1%) of diabetes onset.

Appendix: Scripts Used

setwd("c:/Users/baumgaral/R")

library(readxl)

pima <- read\_excel("c:/Users/baumgaral/Data/pima.xlsx")

attach(pima)

> reg\_mod <- glm(Outcome ~ ., data=pima, family = binomial)

> summary(reg\_mod)

Call:

glm(formula = Outcome ~ ., family = binomial, data = pima)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5566 -0.7274 -0.4159 0.7267 2.9297

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.4046964 0.7166359 -11.728 < 2e-16 \*\*\*

Pregnancies 0.1231823 0.0320776 3.840 0.000123 \*\*\*

Glucose 0.0351637 0.0037087 9.481 < 2e-16 \*\*\*

BP -0.0132955 0.0052336 -2.540 0.011072 \*

SkinThickness 0.0006190 0.0068994 0.090 0.928515

Insulin -0.0011917 0.0009012 -1.322 0.186065

BMI 0.0897010 0.0150876 5.945 2.76e-09 \*\*\*

DiabPedFct 0.9451797 0.2991475 3.160 0.001580 \*\*

Age 0.0148690 0.0093348 1.593 0.111192

---

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: 723.45 on 759 degrees of freedom

AIC: 741.45

Number of Fisher Scoring iterations: 5

> library(car)

> scatterplotMatrix(~Pregnancies + Glucose + BP + BMI + DiabPedFct, data=pima, main="Variables")

> d <- data.frame(Pregnancies,Glucose,BP,BMI,DiabPedFct)

> cor(d)

> log\_mod\_pima <- glm(Outcome~Pregnancies + Glucose + BP + BMI + DiabPedFct, data=pima, family=binomial)

> summary(log\_mod\_pima)

current\_model <- log\_mod\_pima #Run glm(), assign to this object; also update lines 96, 103 and 114

current\_data <- pima #Be sure this points to the current data source

G2 <- current\_model$null.deviance - current\_model$deviance

# EL50 <- -(-12.3508)/0.4972 # Input the EL50 = -a/B (negated intercept / Beta coefficient)

print(summary(current\_model))

writeLines("Deviance Statistic:")

print(G2)

writeLines("")

writeLines("Coefficients Exponentiated:")

print(exp(coef(current\_model)))

writeLines("")

writeLines("")

writeLines("95% CI's (Wald):")

print(confint.default(current\_model))

writeLines("")

writeLines("")

writeLines("95% CI's (Wald) Exponentiated:")

print(exp(confint.default(current\_model)))

writeLines("")

writeLines("")

writeLines("95% CI's (Profile/Likelihood Ratio):")

print(confint(current\_model))

writeLines("")

writeLines("")

writeLines("95% CI's (Profile/Likelihood Ratio) Exponentiated:")

print(exp(confint(current\_model)))

writeLines("")

writeLines("")

pr <- sum(residuals(current\_model, type = "pearson")^2)

df <- current\_model$df.residual

p\_value <- pchisq(pr, current\_model$df.residual, lower=F)

print(matrix(c("Pearson Chi GOF", "Chi2", "df", "p-value", "Parameters", round(pr,4),df,round(p\_value,4)), ncol=2))

writeLines("")

writeLines("")

mu <- current\_model$fitted.value

dr <- resid(current\_model, type = "deviance")

hat <- hatvalues(current\_model)

stdr <- dr/sqrt(1-hat)

windows()

plot(mu, stdr^2)

abline(h = 4, col = "red")

predict <- predict(current\_model)

fit <- current\_model$fitted.values

# Calculate standard errors of the linear predictor

lpred <- predict(current\_model, newdata = current\_data, type = "link", se.fit = TRUE)

up <- lpred$fit + (qnorm(.975) \* lpred$se.fit)

low <- lpred$fit - (qnorm(.975) \* lpred$se.fit)

eta <- lpred$fit

upci <- current\_model$family$linkinv(up)

mu <- current\_model$family$linkinv(eta)

loci <- current\_model$family$linkinv(low)

writeLines("Lower CI:")

print(summary(loci))

writeLines("")

writeLines("Mean CI:")

print(summary(mu))

writeLines("")

writeLines("Upper CI:")

print(summary(upci))

writeLines("")

# Bayseian Information Criterion

library(COUNT)

writeLines("AIC and BIC Statistics:")

print(modelfit(current\_model))

layout(1)

plot(current\_data$BMI, mu, col = 1)

lines(current\_data$BMI, loci, col = 2, type = 'p')

lines(current\_data$BMI, upci, col = 3, type = 'p')

coef <- current\_model$coefficients

se <- sqrt(diag(vcov(current\_model)))

coefse <- data.frame(coef, se)

writeLines("")

pr <- resid(current\_model, type = "pearson")

pchi2 <- sum(residuals(current\_model, type = "pearson")^2)

disp <- pchi2/current\_model$df.residual

scse <- se\*sqrt(disp)

library(sandwich)

rmodel <- glm(current\_data$Outcome ~ current\_data$Pregnancies + current\_data$Glucose + current\_data$BP + current\_data$BMI + current\_data$DiabPedFct, family = binomial, data = current\_data) #Update variables!!

# WITH FACTOR: rmodel <- glm(current\_data$y ~ + current\_data$weight+ current\_data$los + factor(current\_data$type), family = binomial, data = current\_data) #Update variables!!

rse <- sqrt(diag(vcovHC(rmodel, type = "HC0")))

newcoefse <- data.frame( coef, se, scse, rse)

print(newcoefse)

quasibinomialmod <- glm(current\_data$Outcome ~ current\_data$Pregnancies + current\_data$Glucose + current\_data$BP + current\_data$BMI + current\_data$DiabPedFct, family = quasibinomial, data = current\_data) #Update variables!!

# WITH FACTOR: quasibinomialmod <- glm(current\_data$died ~ + current\_data$white + current\_data$hmo + current\_data$los + factor(current\_data$type), family = quasibinomial, data = current\_data) #Update variables!!

writeLines("")

writeLines("")

writeLines("Quasibinomial model:")

print(summary(quasibinomialmod))

1. “Pima People” Wikipedia, accessed at: https://en.wikipedia.org/wiki/Pima\_people [↑](#footnote-ref-1)
2. Ibid. [↑](#footnote-ref-2)
3. Baier, Leslie J., and Robert L. Hanson, “Genetic Studies of the Etiology of Type 2 Diabetes in Pima Indians: Hunting for Pieces to a Complicated Puzzle,” American Diabetes Association, May, 2004. [↑](#footnote-ref-3)