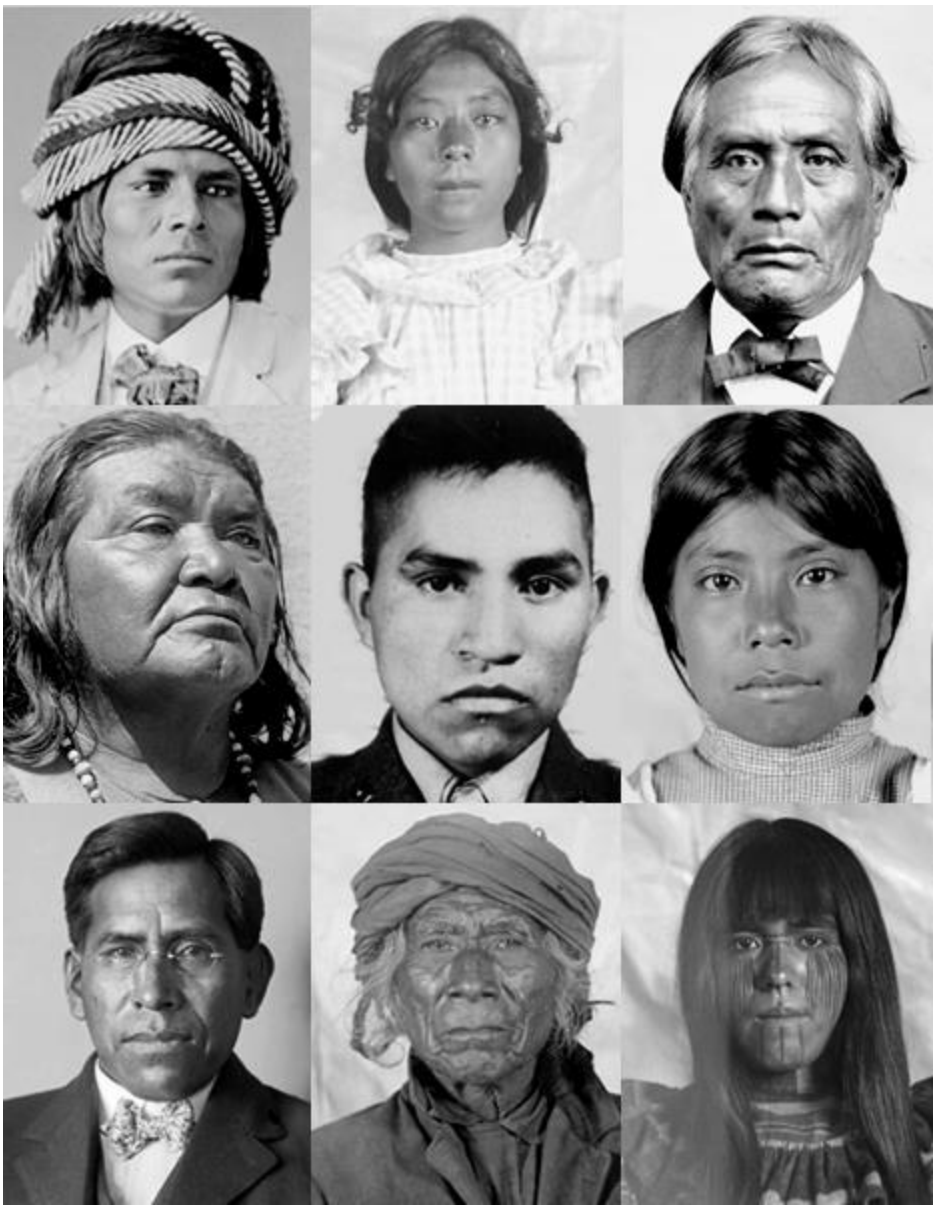


Diabetes in the Pima Indians

PREDICTING DIABETES ONSET

Allen Baumgarten | STAT5120 | 5/3/2018

Overview

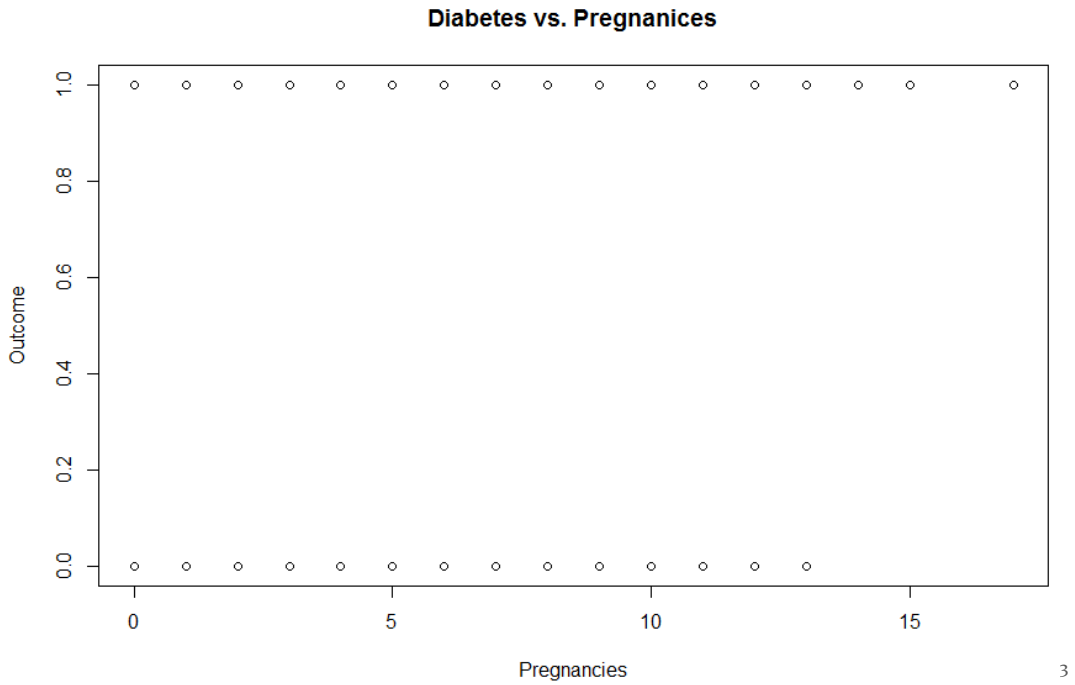
The Pima Indian people are a surviving tribe of native North Americans living in Arizona.[1] Their heritage dates back to at least the late 17th Century if not earlier. Spanish missionaries speak of their tribes and to having made contact with them.[2] Today, the modern descendants of this tribe live 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.

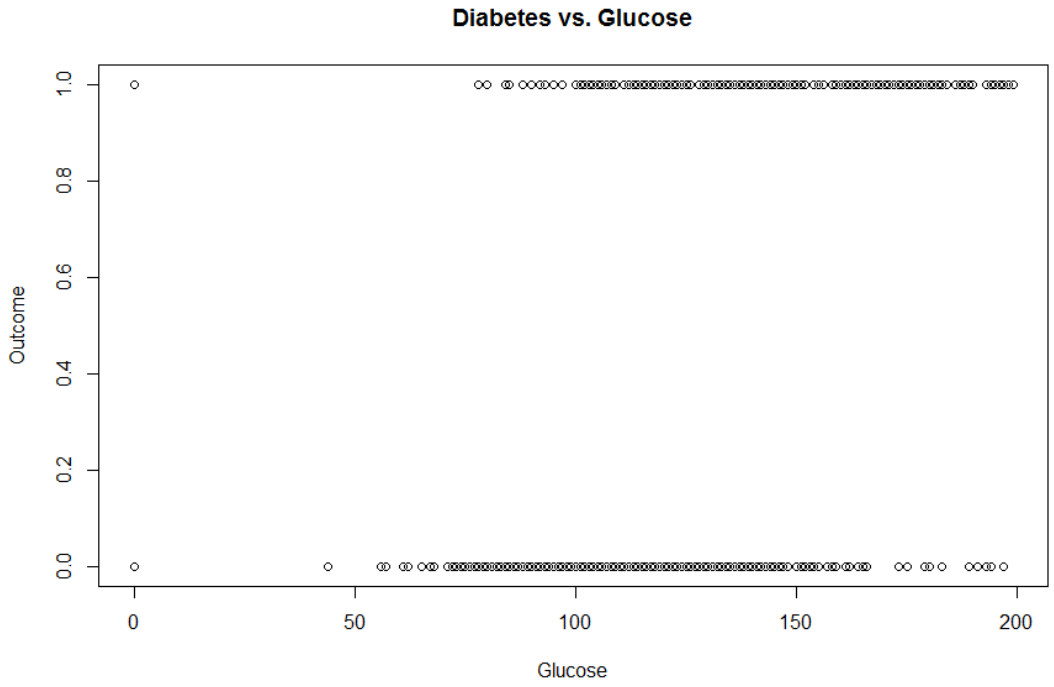


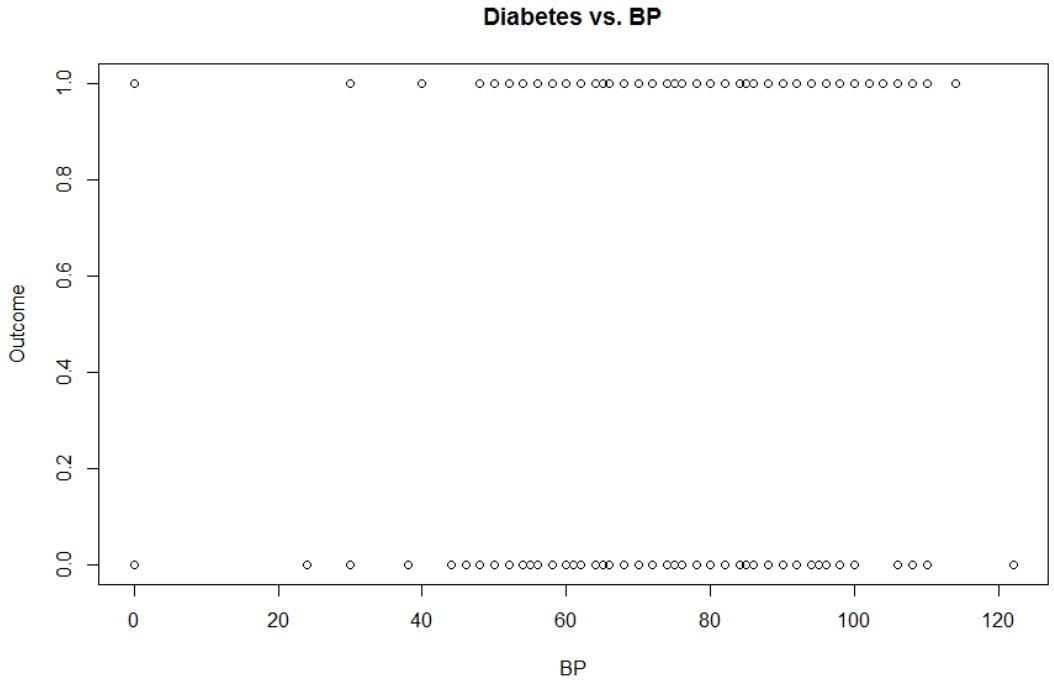
[3]

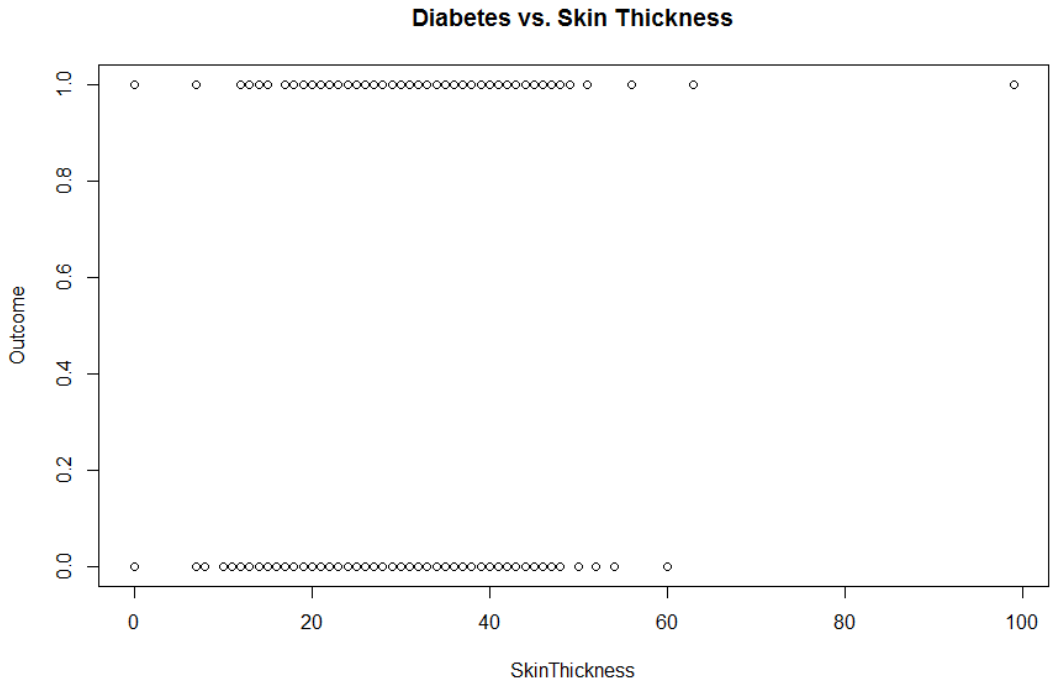
[1] "Pima People" Wikipedia, accessed at: https://en.wikipedia.org/wiki/Pima_people

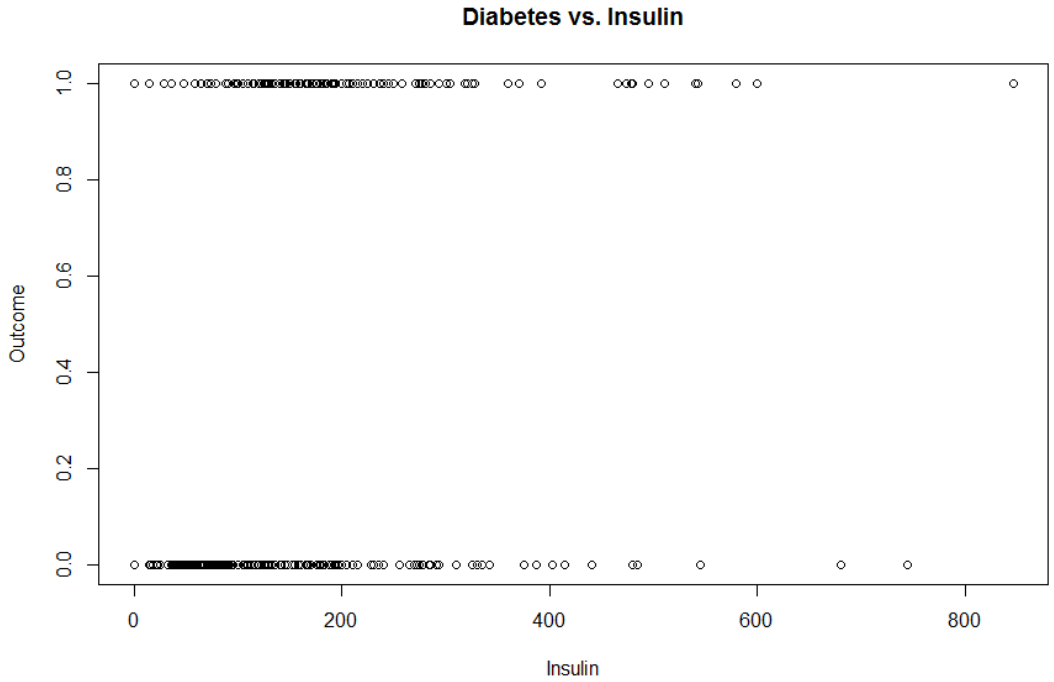
[2] Ibid.

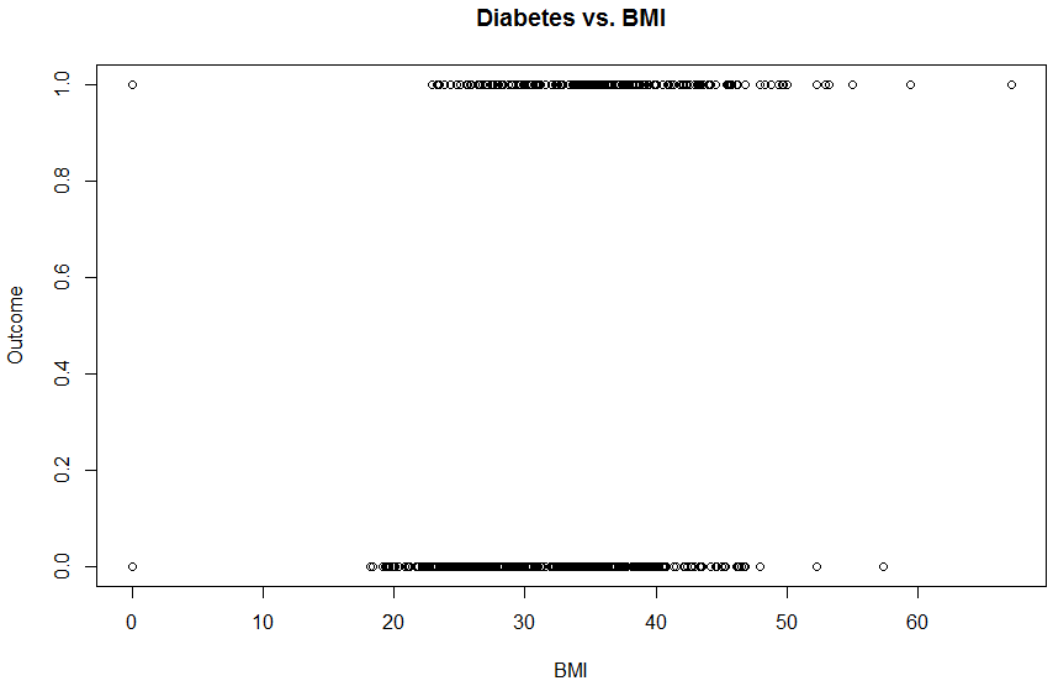
[3] Baier, Leslie J., and Robert L. Hanson, "Genetic Studies of the Etiology of Type 2 Diabetes in Pima Indians: Hunting for Pieces to a Complicated Puzzle," American Diabetes Association, May, 2004.

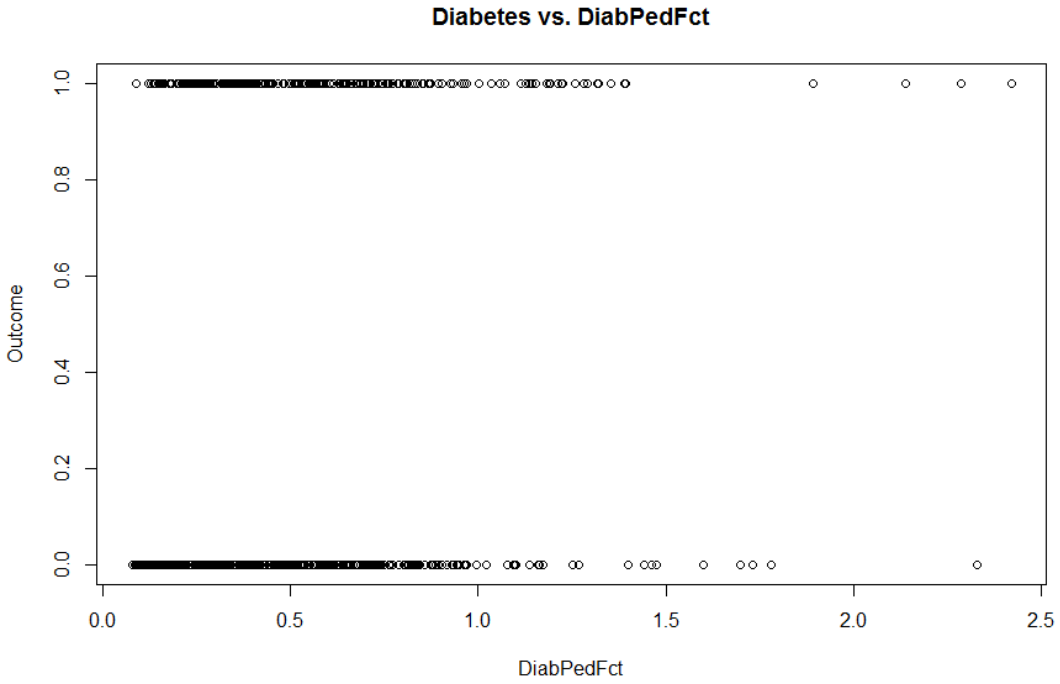

Diabetes vs. Glucose
Outcome
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Glucose

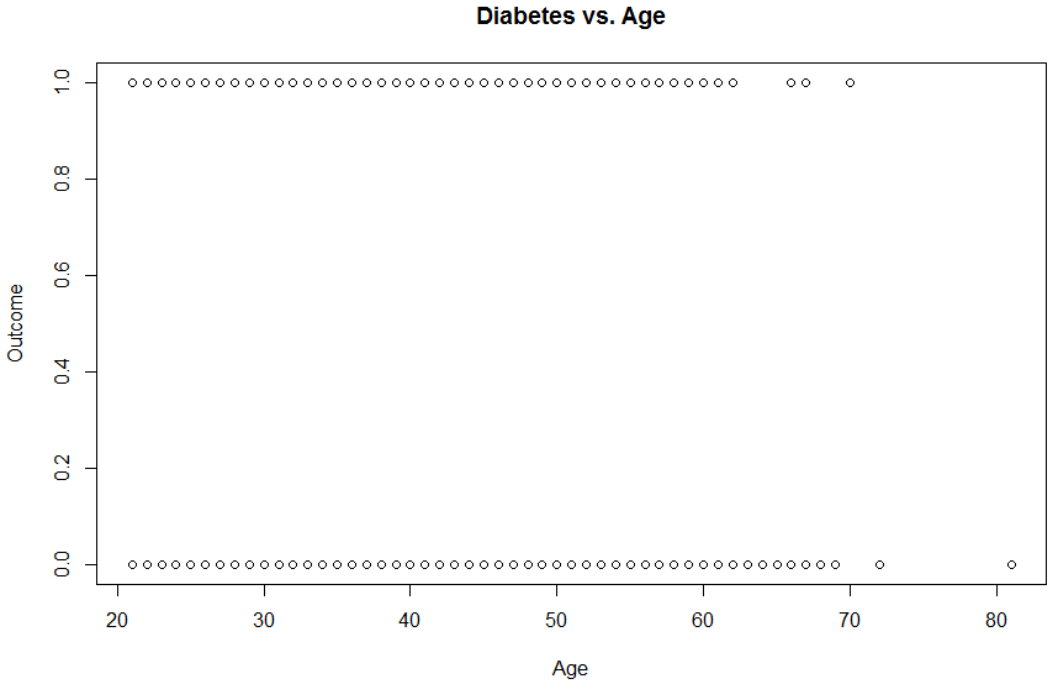

Diabetes vs. BP
Outcome
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
BP


Diabetes vs. Skin Thickness
Outcome
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
SkinThickness


Diabetes vs. Insulin
Outcome
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
Insulin


Diabetes vs. BMI
Outcome
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
BMI


Diabetes vs. DiabPedFct
Outcome
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
DiabPedFct



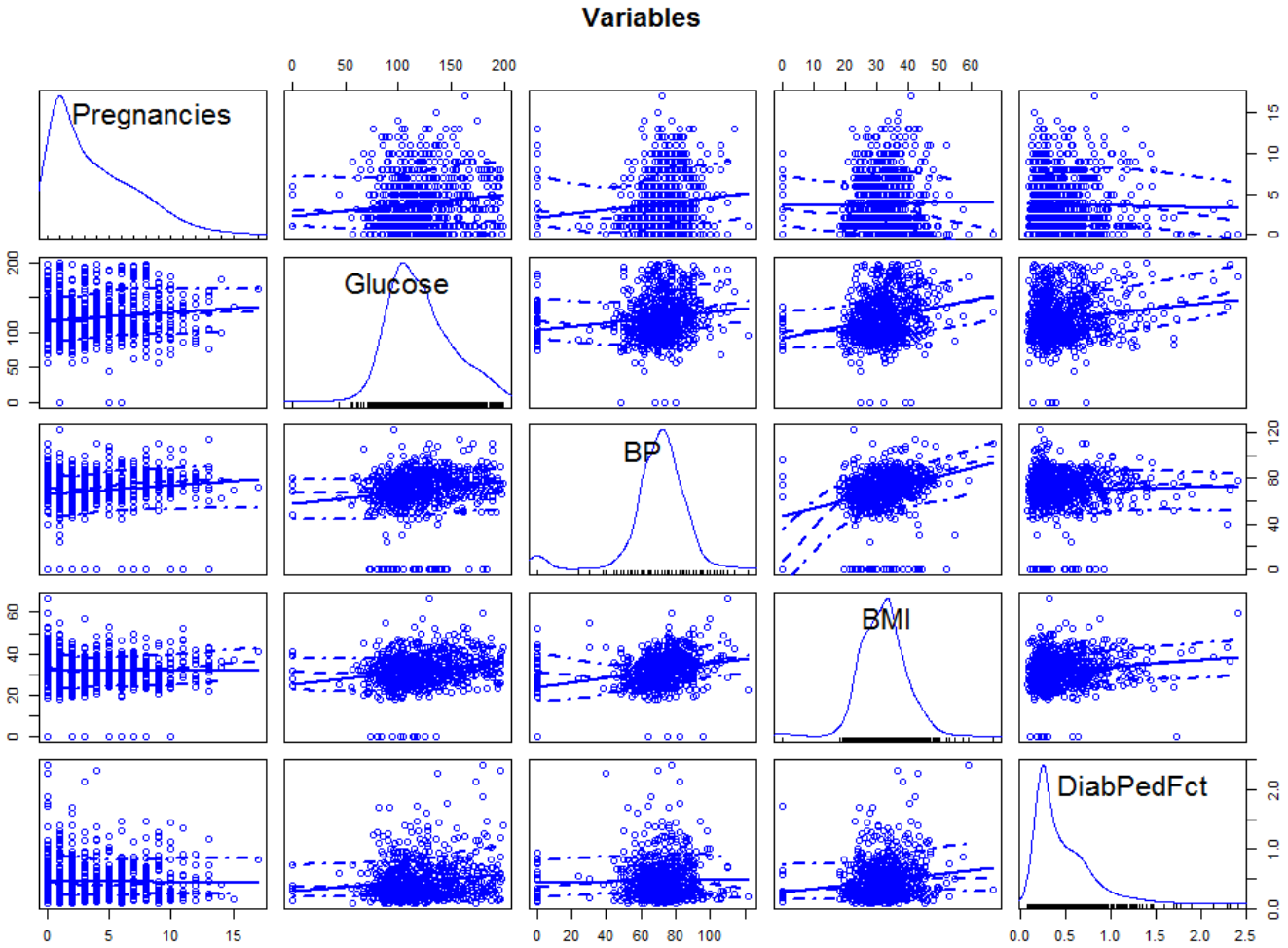
Model Selection Process

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:

```
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
```

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:

	Pregnancies	Glucose	BP	BMI	DiabPedFct
Pregnancies	1.00000000	0.1294587	0.14128198	0.01768309	-0.03352267
Glucose	0.12945867	1.0000000	0.15258959	0.22107107	0.13733730
BP	0.14128198	0.1525896	1.00000000	0.28180529	0.04126495
BMI	0.01768309	0.2210711	0.28180529	1.00000000	0.14064695
DiabPedFct	-0.03352267	0.1373373	0.04126495	0.14064695	1.00000000



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.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7931  -0.7362  -0.4188   0.7251   2.9555

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.954952   0.675823 -11.771  < 2e-16 ***
Pregnancies  0.153492   0.027835   5.514  3.5e-08 ***
Glucose      0.034658   0.003394  10.213  < 2e-16 ***
BP          -0.012007   0.005031  -2.387  0.01700 *
BMI          0.084832   0.014125   6.006  1.9e-09 ***
DiabPedFct   0.910628   0.294027   3.097  0.00195 **
---
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: 728.56  on 762  degrees of freedom
AIC: 740.56

Number of Fisher Scoring iterations: 5
```

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.

```
AIC and BIC Statistics:
$AIC
[1] 740.5596

$AICn
[1] 0.9642703

$BIC
[1] 768.4223

$BICqh
[1] 0.9766416
```

```
Deviance Statistic:
[1] 264.9243

Coefficients Exponentiated:
 (Intercept)  Pregnancies      Glucose           BP          BMI   DiabPedFct
0.0003509201 1.1658987706 1.0352654766 0.9880648446 1.0885341500 2.4858820555


95% CI's (Wald):
                  2.5 %       97.5 %
(Intercept) -9.27954035 -6.630363485
Pregnancies  0.09893608  0.208048453
Glucose      0.02800674  0.041309046
BP          -0.02186662 -0.002147284
BMI          0.05714783  0.112516123
DiabPedFct   0.33434430  1.486910795


95% CI's (Wald) Exponentiated:
                    2.5 %      97.5 %
(Intercept) 9.331401e-05 0.001319683
Pregnancies 1.103996e+00 1.231272826
Glucose     1.028403e+00 1.042174135
BP          9.783707e-01 0.997855019
BMI         1.058812e+00 1.119090300
DiabPedFct  1.397024e+00 4.423409572


95% CI's (Profile/Likelihood Ratio):
Waiting for profiling to be done...
                  2.5 %       97.5 %
(Intercept) -9.32775132 -6.675509710
Pregnancies  0.09956317  0.208834306
Glucose      0.02818247  0.041502468
BP          -0.02197323 -0.002189315
BMI          0.05781646  0.113246356
DiabPedFct   0.33972920  1.493144091


95% CI's (Profile/Likelihood Ratio) Exponentiated:
Waiting for profiling to be done...
                    2.5 %      97.5 %
(Intercept) 8.892197e-05 0.001261429
Pregnancies 1.104688e+00 1.232240806
Glucose     1.028583e+00 1.042375734
BP          9.782664e-01 0.997813080
BMI         1.059521e+00 1.119907794
DiabPedFct  1.404567e+00 4.451068106
```

```
Lower CI:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0008393 0.0864163 0.2041580 0.2860961 0.4362499 0.9622581

Mean CI:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002118 0.121914 0.275832 0.348958 0.536914 0.989946

Upper CI:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.005337 0.172727 0.351161 0.415017 0.638225 0.997377
```

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.

```
     [,1]              [,2]
[1,] "Pearson Chi GOF" "Parameters"
[2,] "Chi2"            "873.2483"
[3,] "df"              "762"
[4,] "p-value"         "0.0031"
```

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.

```
                  coef          se        scse         rse
(Intercept) -7.95495192 0.675822844 0.723476203 0.711305669
Pregnancies  0.15349227 0.027835300 0.029798012 0.028088923
Glucose      0.03465789 0.003393508 0.003632789 0.003808896
BP          -0.01200695 0.005030535 0.005385246 0.004823188
BMI          0.08483197 0.014124825 0.015120789 0.015055784
DiabPedFct   0.91062755 0.294027467 0.314759818 0.345035091


Quasibinomial model:

Call:
glm(formula = current_data$Outcome ~ current_data$Pregnancies +
    current_data$Glucose + current_data$BP + current_data$BMI +
    current_data$DiabPedFct, family = quasibinomial, data = current_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7931  -0.7362  -0.4188   0.7251   2.9555

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -7.954952   0.723476 -10.995  < 2e-16 ***
current_data$Pregnancies  0.153492   0.029798   5.151 3.30e-07 ***
current_data$Glucose      0.034658   0.003633   9.540  < 2e-16 ***
current_data$BP          -0.012007   0.005385  -2.230  0.02606 *
current_data$BMI          0.084832   0.015121   5.610 2.83e-08 ***
current_data$DiabPedFct   0.910628   0.314760   2.893  0.00392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.145995)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 728.56  on 762  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

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:

```
Coefficients Exponentiated:
 (Intercept)  Pregnancies      Glucose           BP          BMI   DiabPedFct
0.0003509201 1.1658987706 1.0352654766 0.9880648446 1.0885341500 2.4858820555
```

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

```
95% CI's (Wald) Exponentiated:
                    2.5 %      97.5 %
(Intercept) 9.331401e-05 0.001319683
Pregnancies 1.103996e+00 1.231272826
Glucose     1.028403e+00 1.042174135
BP          9.783707e-01 0.997855019
BMI         1.058812e+00 1.119090300
DiabPedFct  1.397024e+00 4.423409572
```

```
95% CI's (Profile/Likelihood Ratio) Exponentiated:
Waiting for profiling to be done...
                    2.5 %      97.5 %
(Intercept) 8.892197e-05 0.001261429
Pregnancies 1.104688e+00 1.232240806
Glucose     1.028583e+00 1.042375734
BP          9.782664e-01 0.997813080
BMI         1.059521e+00 1.119907794
DiabPedFct  1.404567e+00 4.451068106
```

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), famil
y = 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 + cur
rent_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))
```