INTRODUCTION
Author: Jacob Haywood

For this project, I chose my data set to be the pima data set from the "faraway" package. In the data set, there are 768 different samples of adult female Pima Indians living near Phoenix. Each column represents a variable that could be linked to whether the person has diabetes or not. The goal of the analysis is to see whether the given variables can predict whether a patient will show signs of diabetes or not in both Frequentist and Bayesian frameworks.

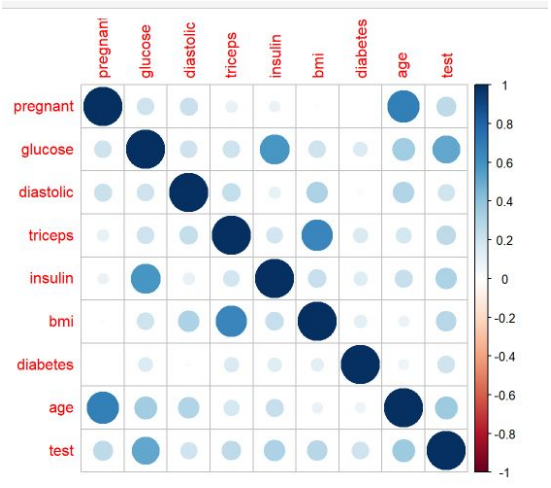
The variables included in the data set are as follows:

pregnant - number of times pregnant
glucose - plasma glucose concentration at 2 hours in an oral glucose tolerance test
diastolic - diastolic blood pressure
triceps - tricep skin fold thickness
insulin - 2 hour serum insulin
bmi - body mass index
diabetes - diabetes pedigree function
age - age
test - whether the patient showed signs of diabetes (1 = Yes, 0 = No)

For the remainder of the analysis, we will refer to each variable by their variable name.

EXPLORATORY DATA ANALYSIS

To begin, we visualize the data using the summary() and plot() functions. At first glance, the information given from these functions seems odd, as many of the insulin values are 0. By setting these 0 values equal to NA and rerunning the summary() and plot() functions, we see a much clearer result. These NA values will then be removed as we continue our analysis.



Next, a correlation chart was plotted to visualize the correlations at first glance. This correlation plot was plotted using the corrplot() function in the “corrplot” package. We can make inferences from this correlation plot. For example, the correlation between pregnancy and test seems to be small, while glucose and

test seem to have a larger correlation. It may be important to notice that no variables at this time have strong negative correlations with test.

METHODS

For this project, we will use two models for our analysis: Frequentist and Bayesian. Our response variable (test) is a binary outcome, so fitting a linear model to test may produce an invalid outcome. This is because an assumption for linear models is that the outcome is continuous, but this is not the case for a binary variable. Thus, we will utilize a logistic model for our outcome.

For our Frequentist approach, we will use the glm() function on our variables and set our family equal to binomial. We will then obtain our coefficient estimates, standard errors, z statistics, and our p-values through the summary() function. Our p-values will then indicate which variables actually have a significant correlation with our response variable. Removal of uncorrelated explanatory variables is key to setting a good model for our correlated explanatory variables, so we will remove the least correlated variable first and run the summary again. We will continue this until we have a model with the best fit. We are looking for the lowest AIC value because a low AIC value implies a better model. Thus, we will continue to remove the variable with the highest p-value until we arrive at our lowest AIC value.

```
fm = glm(test~pregnant+glucose+diastolic+triceps+insulin+bmi+age+diabetes,data=pima_data,family='binomial')
summary(fm) #AIC: 362.14
```

```
Call:
glm(formula = test ~ pregnant + glucose + diastolic + triceps +
    insulin + bmi + age + diabetes, family = "binomial", data = pima_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7742  -0.6593  -0.3615   0.6385   2.5617

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.007e+01  1.215e+00  -8.286  < 2e-16 ***
pregnant     8.253e-02  5.543e-02   1.489  0.13650
glucose      3.829e-02  5.769e-03   6.637 3.21e-11 ***
diastolic   -1.563e-03  1.182e-02  -0.132  0.89476
triceps      1.083e-02  1.702e-02   0.636  0.52482
insulin     -8.299e-04  1.307e-03  -0.635  0.52538
bmi          7.187e-02  2.687e-02   2.675  0.00748 **
age          3.415e-02  1.837e-02   1.859  0.06301 .
diabetes     1.129e+00  4.250e-01   2.658  0.00787 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.70  on 393  degrees of freedom
Residual deviance: 344.14  on 385  degrees of freedom
AIC: 362.14

Number of Fisher Scoring iterations: 5
```

First, we use the summary() function on all variables and we find that diastolic has the highest p-value. In this model, we have an AIC value of 362.14.

```
fm1 = glm(test~pregnant+glucose+triceps+insulin+bmi+age+diabetes,data=pima_data,family='binomial')
summary(fm1) #removed diastolic, AIC: 360.16
```

```
Call:
glm(formula = test ~ pregnant + glucose + triceps + insulin +
    bmi + age + diabetes, family = "binomial", data = pima_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7748  -0.6628  -0.3596   0.6347   2.5716

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.013e+01  1.129e+00  -8.967  < 2e-16 ***
pregnant     8.234e-02  5.536e-02   1.487  0.13696
glucose      3.821e-02  5.731e-03   6.666 2.63e-11 ***
triceps      1.079e-02  1.702e-02   0.634  0.52609
insulin     -8.177e-04  1.303e-03  -0.628  0.53027
bmi          7.102e-02  2.612e-02   2.719  0.00654 **
age          3.369e-02  1.801e-02   1.871  0.06138 .
diabetes     1.132e+00  4.243e-01   2.669  0.00762 **
```

After the removal of diastolic, we see that more variables now have significant p-values. We

also get an AIC value of 360.16, which is an improvement. Again, we repeat this process.

```
fm2 = glm(test~pregnant+glucose+triceps+bmi+age+diabetes,data=pima_data,family='binomial')
summary(fm2) #removed insulin, AIC: 358.55
```

```
Call:
glm(formula = test ~ pregnant + glucose + triceps + bmi + age +
    diabetes, family = "binomial", data = pima_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8868  -0.6542  -0.3588   0.6326   2.5593

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.947393   1.084898  -9.169  < 2e-16 ***
pregnant     0.083928   0.055160   1.522  0.12813
glucose      0.036488   0.004988   7.315 2.57e-13 ***
triceps      0.011173   0.016993   0.657  0.51087
bmi          0.068339   0.025690   2.660  0.00781 **
age          0.033055   0.017970   1.839  0.06585 .
diabetes     1.119098   0.423245   2.644  0.00819 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.70  on 393  degrees of freedom
Residual deviance: 344.55  on 387  degrees of freedom
AIC: 358.55

Number of Fisher Scoring iterations: 5
```

After the removal of the variable with the next highest p-value, insulin, we see that the AIC value dropped to 358.55 which is a slight improvement. Thus, we remove our next highest p-value, tricep.

```
fm3 = glm(test~pregnant+glucose+bmi+age+diabetes,data=pima_data,family='binomial')
summary(fm3) #removed tricep, AIC: 356.99
```

```
Call:
glm(formula = test ~ pregnant + glucose + bmi + age + diabetes,
    family = "binomial", data = pima_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8772  -0.6514  -0.3641   0.6483   2.5819

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.016534   1.083354  -9.246  < 2e-16 ***
pregnant      0.084233   0.055037   1.530 0.125896
glucose       0.036462   0.004978   7.325 2.38e-13 ***
bmi           0.078848   0.020399   3.865 0.000111 ***
age           0.034447   0.017809   1.934 0.053086 .
diabetes      1.141246   0.422040   2.704 0.006849 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.70  on 393  degrees of freedom
Residual deviance: 344.99  on 388  degrees of freedom
AIC: 356.99

Number of Fisher Scoring iterations: 5
```

Removing the tricep variable decreased our AIC value to 356.99. We repeat this process again with our pregnant variable.

```
fm4 = glm(test~glucose+bmi+age+diabetes,data=pima_data,family='binomial')
summary(fm4) #removed pregnant, AIC: 357.35
```

```
Call:
glm(formula = test ~ glucose + bmi + age + diabetes, family = "binomial",
    data = pima_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8248  -0.6592  -0.3739   0.6621   2.5887

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.119951   1.076257  -9.403  < 2e-16 ***
glucose       0.036194   0.004982   7.265 3.73e-13 ***
bmi           0.075236   0.020042   3.754 0.000174 ***
age           0.053179   0.013429   3.960 7.49e-05 ***
diabetes      1.075759   0.416731   2.581 0.009839 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.70  on 393  degrees of freedom
```

The removal of pregnant **increased** our AIC value to 357.35, so we can conclude that the removal of pregnant did not return the best model. Thus, our final variables in our model are

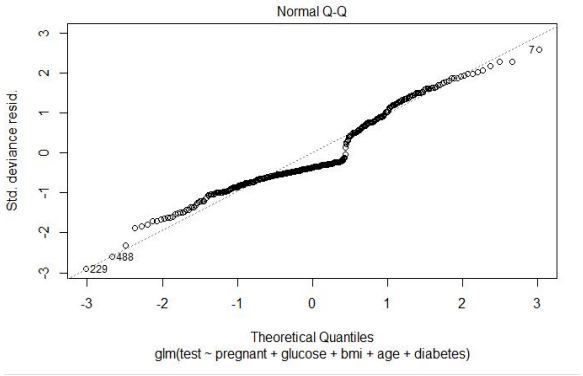
pregnant, glucose, bmi, age, and diabetes.

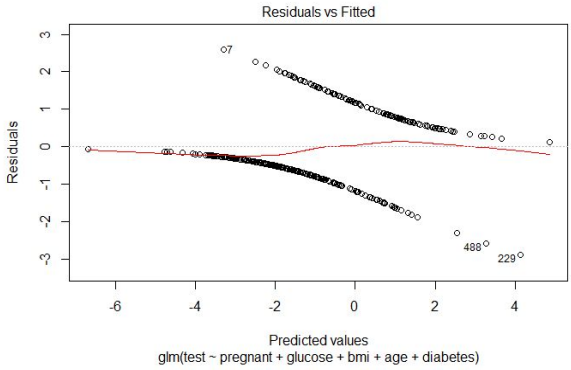
Our final model will be the following:

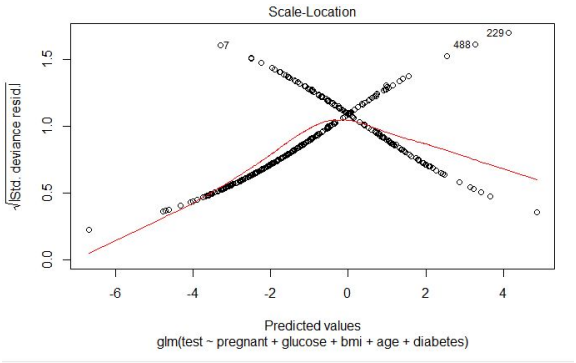
$$log\left(\frac{\theta}{1-\theta}\right) = -10.0165 + 0.0842 \cdot pregnant + 0.0365 \cdot glucose + 0.0788 \cdot bmi + 0.0344 \cdot age + 1.1412 \cdot diabetes$$

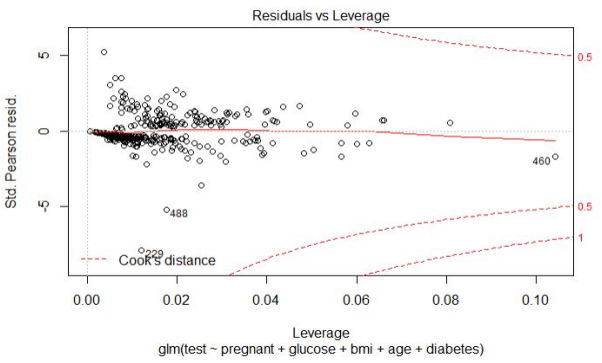
Now that we have our final model, we can plug our parameters into the model to predict theta, the probability of testing for diabetes.

These are our residuals and Q-Q plot for our frequentist approach. A Q-Q plot is utilized to see how normally distributed our residuals are. Our Q-Q plot shows slight skews at both ends, and a large skew in the middle. It may be important to notice that because our data does not seem very normally distributed, a Bayesian analysis with a normal prior may not be very useful.





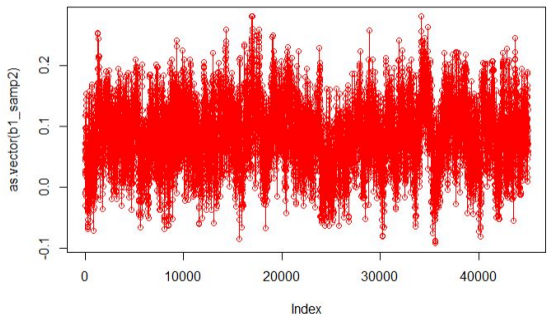




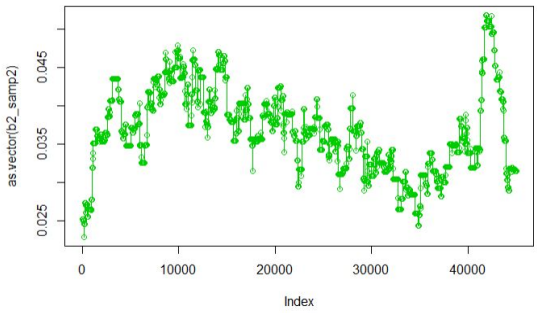
For our bayesian approach, a function named logisticRegressionBayes() was written. This will allow us to utilize a Metropolis algorithm to generate 50,000 samples from the posterior distribution of the logistic regression. After running the logisticRegressionBayes() function, we can run the summary() function to see some useful information. The bottom also gives the credible intervals for each of our variables as well, with a 95% interval being from the 2.5% value to the 97.5% value.

```
Iterations = 1:50000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 50000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean       SD  Naive SE
(Intercept) -10.26006 1.529437 6.840e-03
pregnant      0.08558 0.051985 2.325e-04
glucose       0.03479 0.007968 3.563e-05
bmi           0.08957 0.019658 8.791e-05
age           0.03703 0.016806 7.516e-05
diabetes      1.13947 0.418230 1.870e-03
          Time-series SE
(Intercept)       0.267642
pregnant          0.002314
glucose           0.002482
bmi               0.002773
age               0.002131
diabetes          0.015309

2. Quantiles for each variable:

                 2.5%       25%       50%       75%
(Intercept) -12.762681 -11.16822 -10.40703 -9.60328
pregnant     -0.015384   0.05095   0.08608  0.12022
glucose       0.014110   0.03141   0.03561  0.03974
bmi           0.049451   0.07727   0.08961  0.10242
age           0.003514   0.02534   0.03761  0.04855
diabetes      0.350793   0.85098   1.13014  1.41775
                97.5%
(Intercept) -6.52140
pregnant     0.18825
glucose      0.04700
bmi          0.12810
age          0.06904
diabetes     1.98006
```

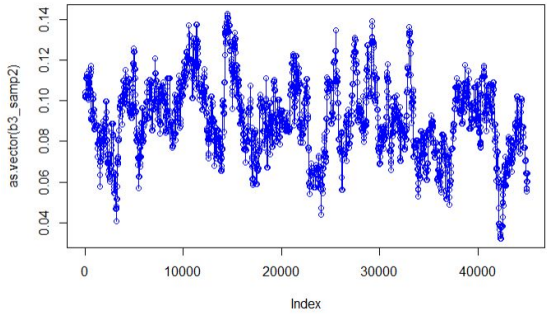
We will then plot trace plots for each of our explanatory variables. Then, we want to target our trace plot at its convergence so we will utilize burn-ins to remove initial fluctuations. After utilizing a burn in of 5000, we get these trace plots.



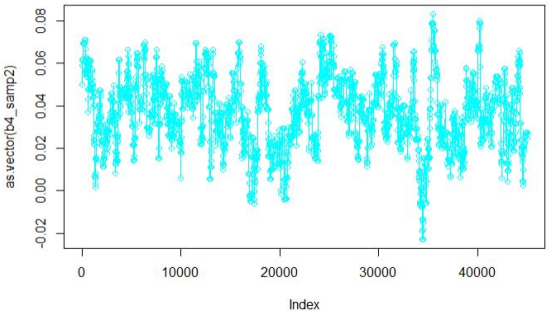
Pregnant



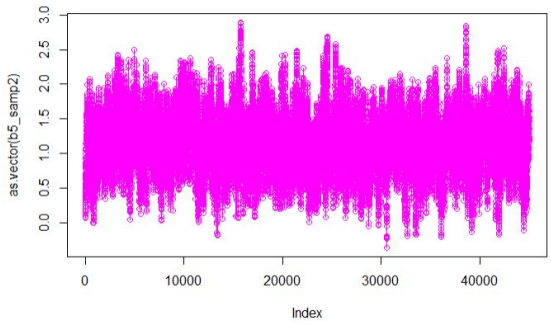
Glucose



BMI



Age



Diabetes

MODEL DIAGNOSIS AND VARIABLE SELECTION

Now we should decide what variance option to use. First, we run the function for 3 values for the variance (0.01, 0.001, 0.0001) and find the acceptance rate and effective size.

This is a table for the values:

Variance	Acceptance Rate	Effective Size
0.01	0.345	(Intercept) pregnant glucose bmi age diabetes 38.57315 321.74894 35.41846 52.71640 67.47621 431.60552
0.001	0.479	(Intercept) pregnant glucose bmi age diabetes 14.88493 605.57461 45.46014 84.16676 182.93834 67.02963
0.0001	0.628	(Intercept) pregnant glucose bmi age diabetes 2.448726 278.126195 63.867248 38.715259 162.902586 7.048951

Lag-50 Correlation
, , (Intercept) (Intercept) pregnant glucose bmi age diabetes Lag 50 0.9313246 0.1059985 -0.7163429 -0.7241788 -0.3410547 -0.2181099 , , pregnant (Intercept) pregnant glucose bmi age diabetes Lag 50 0.09185663 0.457439 0.05345864 0.1461687 -0.5983484 0.1008317 , , glucose (Intercept) pregnant glucose bmi age diabetes Lag 50 -0.7059312 0.05199099 0.9294444 0.3816911 0.01241978 0.1782745 , , bmi (Intercept) pregnant glucose bmi age diabetes Lag 50 -0.7398505 0.1180967 0.3863595 0.9132262 0.03793329 0.08433999 , , age (Intercept) pregnant glucose bmi age diabetes Lag 50 -0.3352964 -0.6073625 0.01461445 0.01094489 0.8738428 -0.01953599 , , diabetes (Intercept) pregnant glucose bmi age diabetes Lag 50 -0.2058912 0.1370075 0.2196273 0.05739146 -0.07174793 0.4433628

```
, , (Intercept)

       (Intercept)   pregnant    glucose        bmi        age   diabetes
Lag 50    0.970659 0.06644216 -0.7168787 -0.6822751 -0.3526302 -0.3672102

, , pregnant

       (Intercept) pregnant    glucose       bmi        age  diabetes
Lag 50  0.05933735  0.33375 0.01111076 0.1121833 -0.4410749 0.0660394

, , glucose

       (Intercept)    pregnant   glucose       bmi        age  diabetes
Lag 50  -0.7179917 -0.01457624 0.9144028 0.2348953 0.09187737 0.2598258

, , bmi

       (Intercept)  pregnant   glucose       bmi       age  diabetes
Lag 50  -0.6850414 0.1131504 0.2289192 0.8465161 0.1414138 0.1176685

, , age

       (Intercept)   pregnant    glucose      bmi       age   diabetes
Lag 50  -0.3442057 -0.4393232 0.07718999 0.125018 0.6935456 0.02832899

, , diabetes

       (Intercept)   pregnant   glucose       bmi          age  diabetes
Lag 50  -0.3697962 0.09274097 0.2841913 0.1185294 -0.008091756 0.8706591
```

```
, , (Intercept)

       (Intercept) pregnant    glucose        bmi        age  diabetes
Lag 50   0.9950922 0.126296 -0.8303628 -0.9062783 -0.6059528 -0.795208

, , pregnant

       (Intercept) pregnant     glucose         bmi       age    diabetes
Lag 50   0.1219486 0.592223 -0.02083469 -0.05063041 -0.482931 -0.03573972

, , glucose

       (Intercept)    pregnant   glucose       bmi       age  diabetes
Lag 50   -0.834689 -0.04711482 0.8931983 0.6697007 0.3815213 0.6343785

, , bmi

       (Intercept)    pregnant   glucose       bmi       age  diabetes
Lag 50   -0.904218 -0.03171885 0.6595225 0.9055374 0.5298114 0.6969512

, , age

       (Intercept)   pregnant   glucose       bmi       age  diabetes
Lag 50  -0.6027313 -0.4932504 0.3777695 0.5276497 0.7197797 0.4258014

, , diabetes

       (Intercept)    pregnant   glucose       bmi       age  diabetes
Lag 50  -0.7933108 -0.04994302 0.6245191 0.6973398 0.4368973 0.9854982
```

After analyzing the table, we decide to go with a variance of 0.01 because of its higher effective size on average and smallest average lag-50 correlation across the variables. It is important to pick the right variance value because very small variances will lead to high rates of acceptance, which is great for the model, but it also leads to high correlation between samples, which may disrupt our model. A balance between the two is crucial; thus, a variance of 0.01 will be optimal.

For our Bayesian model, we should also discuss our likelihood function, prior distribution, and posterior distribution. The likelihood can be found using our original equation that we found through our Frequentist model:

$$log\left(\frac{\theta}{1-\theta}\right) = -10.0165 + 0.0842 \cdot pregnant + 0.0365 \cdot glucose + 0.0788 \cdot bmi + 0.0344 \cdot age + 1.1412 \cdot diabetes$$

By solving for theta, we get:

$$\theta = \left(\frac{\left(e^{-10.0165+0.0842\cdot\ pregnant+0.0365\cdot\ glucose+0.0788\cdot\ bmi+0.0344\cdot\ age+1.1412\cdot\ diabetes}\right)}{1+e^{-10.0165+0.0842\cdot\ pregnant+0.0365\cdot\ glucose+0.0788\cdot\ bmi+0.0344\cdot\ age+1.1412\cdot\ diabetes}} \right)$$

And because our outcome follows a binomial distribution:

$$Likelihood_i\ =\ \Theta_i^{yi}\cdot(1-\Theta_i)^{1-y_i}$$

Thus our final likelihood function is:

$$\theta = \prod_{i=1}^{n} \left(\frac{\left(e^{-10.0165+0.0842\cdot\ pregnant_i+0.0365\cdot\ glucose_i+0.0788\cdot\ bmi_i+0.0344\cdot\ age_i+1.1412\cdot\ diabetes_i}\right)}{1+e^{-10.0165+0.0842\cdot\ pregnant_i+0.0365\cdot\ glucose_i+0.0788\cdot\ bmi_i+0.0344\cdot\ age_i+1.1412\cdot\ diabetes_i}} \right)^{y_i}$$
$$(1-\left(\frac{\left(e^{-10.0165+0.0842\cdot\ pregnant_i+0.0365\cdot\ glucose_i+0.0788\cdot\ bmi_i+0.0344\cdot\ age_i+1.1412\cdot\ diabetes_i}\right)}{1+e^{-10.0165+0.0842\cdot\ pregnant_i+0.0365\cdot\ glucose_i+0.0788\cdot\ bmi_i+0.0344\cdot\ age_i+1.1412\cdot\ diabetes_i}} \right))^{1-y_i}$$

This function can be multiplied to the prior distribution to find our posterior. In this case, we used an uninformative prior because our Q-Q plot showed that a normal prior would not be very accurate.

CONCLUSION

In conclusion, we developed a model for both Frequentist and Bayesian frameworks to predict whether a person has diabetes or not given the other explanatory variables. One thing to notice is that Frequentist models are dependent on large sample sizes and Bayesian models are dependent on large numbers of iterations.This was solved by using 50,000 iterations for our Bayesian model, though including more iterations would improve our accuracy. While there are advantages and disadvantages for both Frequentist and Bayesian models, our Frequentist model may be advantageous due to our model having no useful prior information. Our Bayesian model, however, could be useful for future endeavors due to its ability to incorporate prior information and extend to more levels.

Citations

"Bayesian Inference for Logistic Regression Parameters." *McGill*, www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-621/bayeslogit.pdf.