From Data to Diagnosis: Exploring Diabetes Predictors through Bootstrap, Jackknife, and Inference Models

Afundar, Audrie Lex L.
Cuerdo, Naomi Hannah A.
Rodillas, Christian Miguel T.

In partial fulfillment of the requirements for:
APM1210 - Statistical Computing

Dr. May Anne Tirado

July 20, 2025

Introduction

Diabetes mellitus is a chronic and potentially life-threatening condition that affects millions of people worldwide. It is characterized by the body's inability to produce or effectively use insulin, which leads to elevated blood glucose levels. Early detection of diabetes is crucial to preventing serious health complications such as heart disease, kidney failure, and neuropathy.

The rise of data-driven healthcare has enabled researchers to develop predictive models that can assist in early diagnosis. With the help of research techniques and inferential statistics, this project aims to explore the relationships between medical indicators and diabetes risk to develop more robust, accurate, and interpretable insights.

This study uses the Pima Indians Diabetes Dataset, a well-known dataset in the medical and machine learning communities, in order to investigate which clinical features are most informative for predicting the onset of diabetes.

Statement to the Problem

Identifying the most significant predictors of diabetes remains a challenge due to variable distribution, sampling bias, and model overfitting. **Thus, this project aims to determine which health-related variables most significantly predict the presence of diabetes in patiences, using statistical methods such as bootstrapping, jackknife resampling, permutation testing, and Bayesian inference.**

Particularly, it will try to answer the following questions:

1. What is the main driving factor for the underlying risk of diabetes that can be identified among pregnant women and how much does it affect?

2. Do the distributions of key variables align with known theoretical distributions such as the normal or binomial distribution?

3. What is the main driving factor among clinical indicators that most significantly predicts the risk of diabetes in pregnant women, and to what extent does it affect the probability of developing diabetes?

Exploratory Data Analysis

To better understand the structure and quality of the dataset, it is best to conduct a simple exploratory data analysis, to ensure the nature of each variable is inspected before applying statistical inference techniques in later components.

This shows the head() of the entire dataset:

Pregnancies	**Glucose**	**BloodPressure**	**SkinThickness**	**Insulin**	**BMI**	**DiabetesPedigreeFunction**	**Age**	**Outcome**
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	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	0	0	25.6	0.201	30	0

Table 1. Overview of the Dataset

The dataset contains 768 observations and 9 variables, including 8 predictor features and 1 binary outcome variable (Outcome), where:

- 0 indicates a non-diabetic patient
- 1 indicates a diabetic patient

Variable Summary

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Pregnancies	0	1	3	3.85	6	17
Glucose	0	99	117	120.90	140.2	199
Blood Pressure	0	62	72	69.11	80	122
Skin Thickness	0	0	23	20.54	32	99
Insulin	0	0	30.5	79.80	127.2	846
BMI	0	27.30	32.00	31.99	36.60	67.10
Diabetes Pedigree Function	0.078	0.2437	0.3725	0.4719	0.6262	2.42
Age	21	24	29	33.24	41	81
Outcome	0	0	0	0.349	1	

Table 2. Variable Summary

From the summary, the following findings are observed:

- Glucose, BMI, and Age have reasonable ranges, but may not be normally distributed.
- Insulin and SkinThickness contain many zeros, which may represent missing values rather than true measurements.
- Outcome is binary and imbalanced (approx. 65% non-diabetic, 35% diabetic).

Scope and Limitations

This study aims to explore and identify the most significant clinical predictors of diabetes using the Pima Indians Diabetes Dataset, which includes performing exploratory data analysis, visualizing variable distributions, and applying statistical inference techniques such as bootstrap, jackknife, permutation testing, and Bayesian inference. The dataset provides eight health-related features and one binary outcome. All analysis will be conducted using R.

However, several limitations are needed to be acknowledged. Several features contain zero values that likely represent missing data. The dataset also only includes female patients of Pima Indian descent, limiting the generalization of findings to broader populations. The dataset is cross-sectional, lacking any temporal information (e.g., disease progression or patient follow-up).

The analysis is limited to basic to intermediate statistical methods due to the academic scope of the course (no advanced ML techniques).
Thus, it is important to interpret these results within the context of these limitations.

Probability Density Estimation

Understanding the underlying distribution of medical variables is essential in statistical modeling and inference. In this component, performing Probability Density Estimation on key continuous variables in the dataset, including Glucose, BMI, age, and Insulin. This will allow us to assess distribution, shape, skewness, and compare them with theoretical distributions.

Furthermore, this component explores the empirical distributions of key variables and compares them to theoretical probability distributions, namely normal and binomial distribution, to guide appropriate model selection and transformations for predicting which variable predicts diabetes.

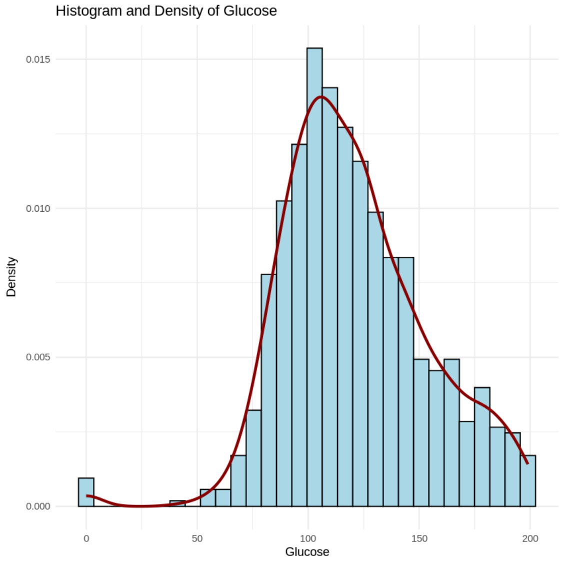


Figure 1. Histogram and Density for Glucose

Figure 1 shows the histogram and KDE plot of Glucose. From the plot, it shows that glucose levels have a right-skewed distribution, indicating that the majority of individuals in the dataset have glucose values clustered between 90 and 130 mg/dL, which aligns with expected normal or prediabetic ranges.

The peak density occurs around 100 mg/dL, suggesting that this is the most common glucose level in the sample. The long tail on the right indicates the presence of individuals with significantly higher glucose levels, possibly diabetic or undiagnosed cases.

The smooth density curve (KDE) overlays the histogram and clearly demonstrates the non-symmetric shape. This right-skewness suggests that the assumption of normality does not hold for this variable.

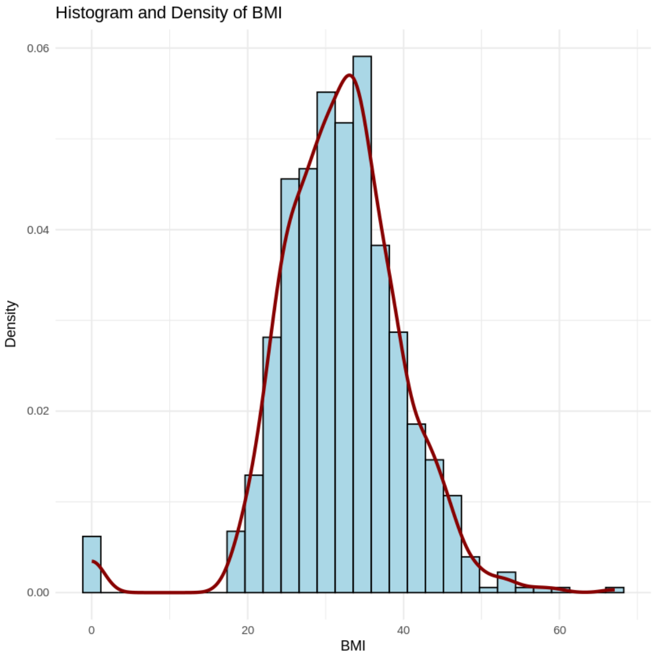


Figure 2. Histogram and Density for BMI

Figure 2 shows the histogram and density for Body Mass Index (BMI) from the plot, it seems that the distribution appears to be approximately normal, but it is slightly right-skewed. Most BMI values fall between 25 and 40, suggesting a population with a high prevalence of overweight and obese individuals. The KDE curve is smooth and unimodal, and the shape is almost symmetric. This makes BMI a good candidate for parametric analyses.

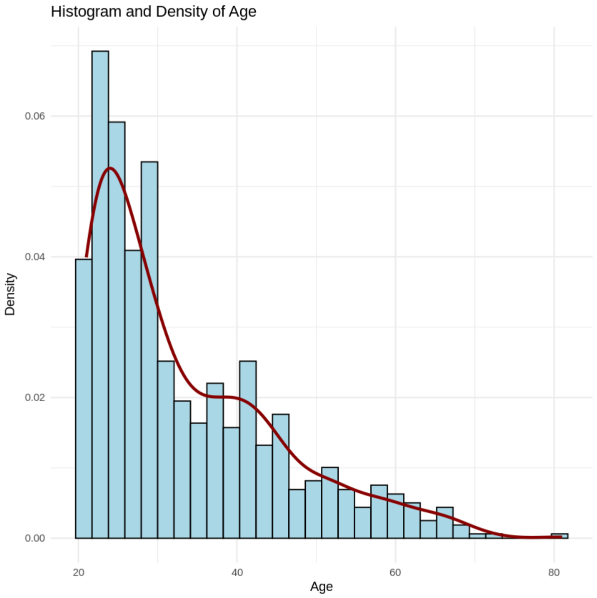


Figure 3. Histogram and Density for Age

The histogram shown above shows a somewhat bimodal distribution, which peaks in both younger and older age brackets. This indicates the presence of two distinct groups in the dataset— possibly younger individuals being screened for possible risks, and older individuals already showing symptoms.

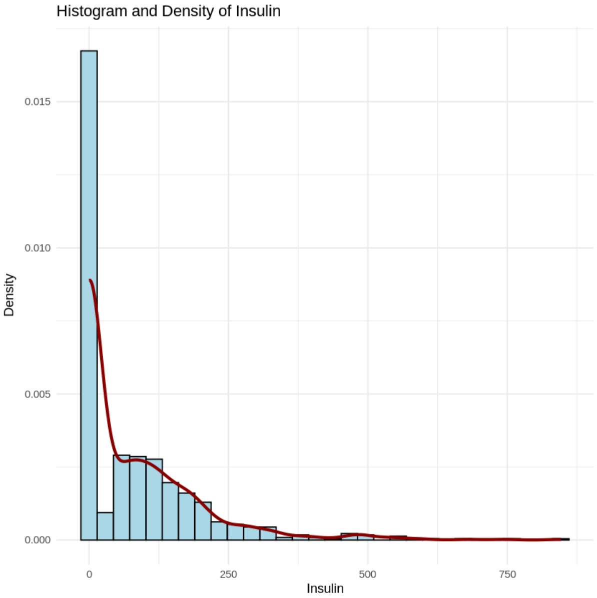


Figure 4. Histogram and Density for Insulin

Figure 4 shows that the insulin distribution is highly skewed to the right, with a large spike near zero. The remaining values are spread out with a long tail, showing significant variability. This kind of distribution is problematic for parametric models and might require data cleaning (e.g., treating zeros as missing) or transformation (e.g., log transformation) before being used in statistical inference.

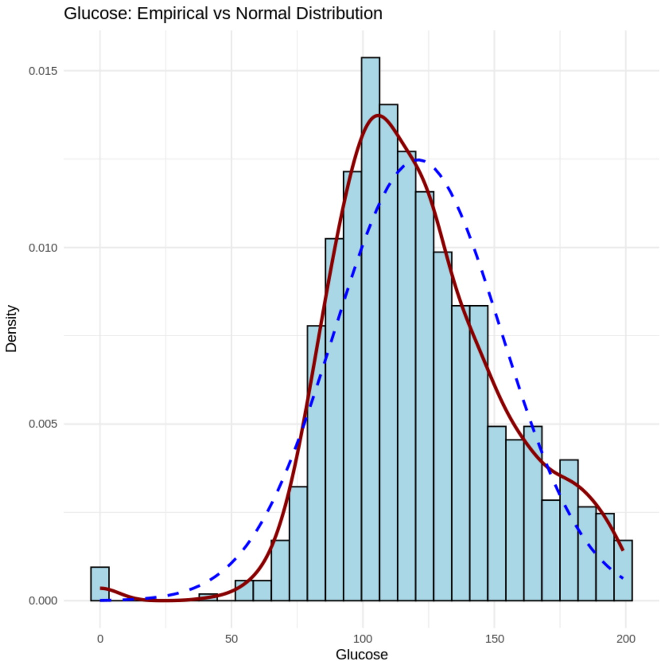


Figure 5. Empirical vs. Normal Distribution of Glucose

Figure 5 shows a comparison of empirical and normal distribution of Glucose. From the figure, it shows that glucose does not follow a normal distribution, which suggests a need for transformation or non-parametric modeling.

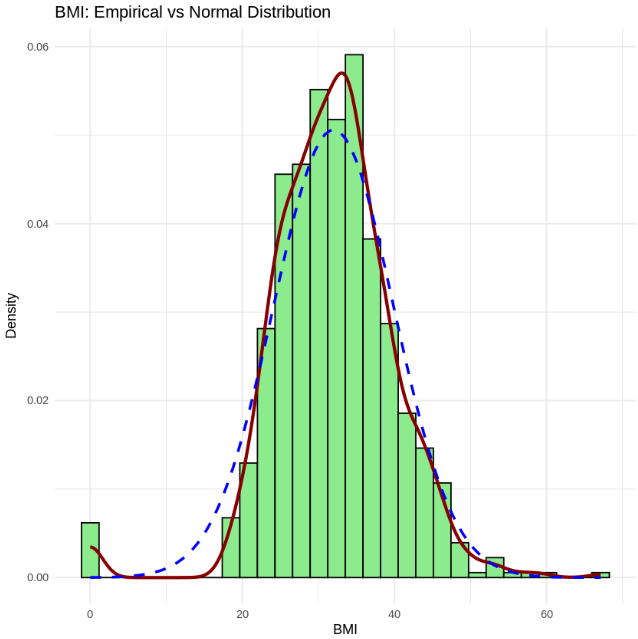


Figure 6. Empirical vs. Normal Distribution of BMI

Figure 6 shows a comparison of empirical and normal distribution of BMI. From the figure, it shows that BMI closely approximates a normal distribution, in which parametric tests are appropriate for this variable.

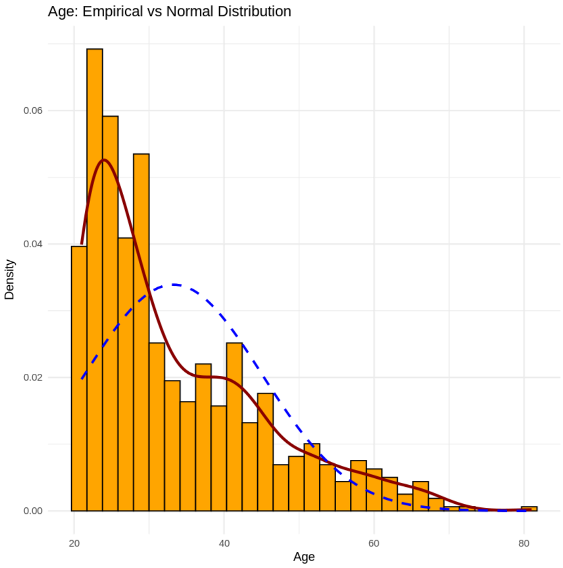


Figure 7. Age vs Normal Distribution

Figure 7 shows the comparison of empirical and normal distribution of Age. From the figure, it shows that age is not normally distributed, in which stratified analysis may be appropriate for this variable.

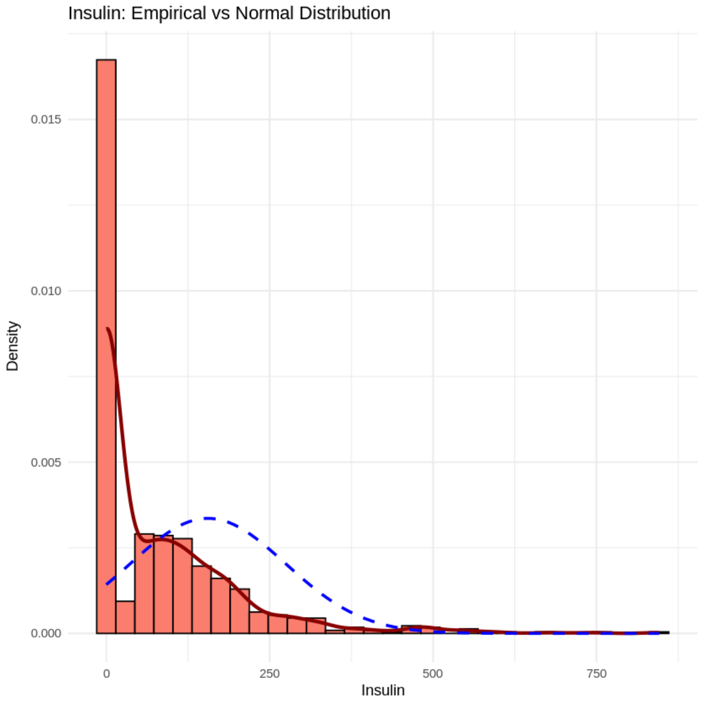


Figure 8. Insulin vs Normal Distribution

Figure 8 shows the comparison between the empirical and normal distribution of insulin. From the plot, it shows that insulin is highly skewed, and it needs cleaning transformation before conducting such analysis.

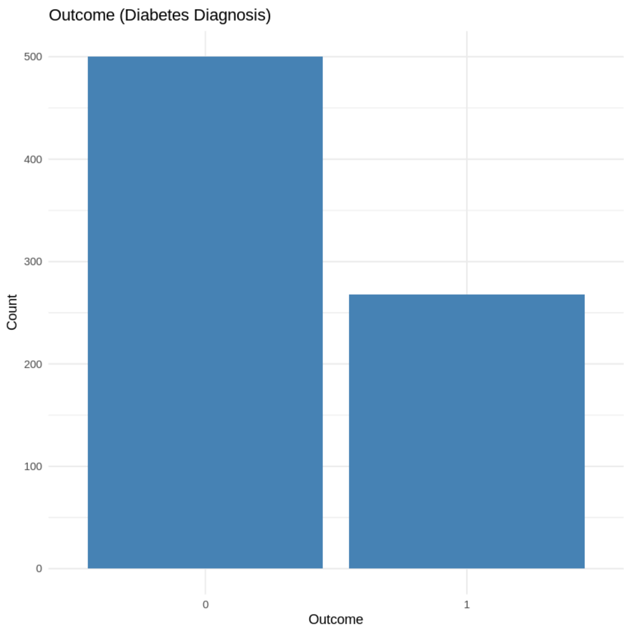


Figure 9. Outcome of Diagnostic Diagnosis

Figure 9 shows a bar graph of the outcome of a patient, whether a patient is diabetic (1) or non-diabetic (0). The bar plot above shows that approximately 500 individuals are non-diabetic, while 268 are diabetic, which aligns with a binomial distribution with unequal probabilities.

This confirms that logistic regression is an appropriate modeling technique for predicting diabetes based on other health-related factors in the dataset, since the binary nature of this variable makes it suitable for classification models.

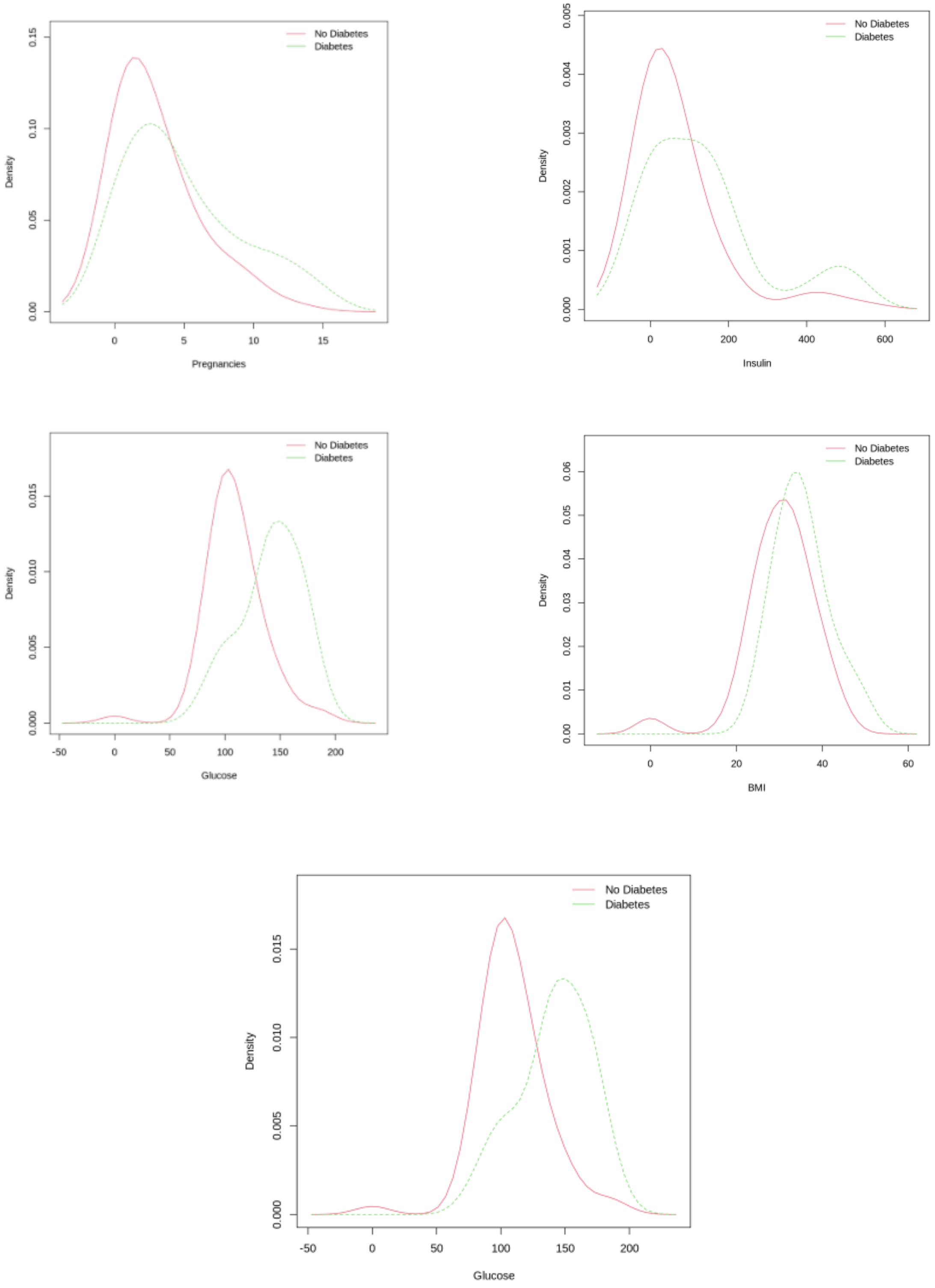
Statistical Testing

For experimenting and learning purposes, the researchers used only 100 of the dataset with stratified sampling to retain the shape of the dataset itself. With this, we can effectively use empirical p-values to determine whether we can reject or accept the null hypothesis.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	103	30	38	83	43.3	0.183	33	0
3	88	58	11	54	24.8	0.267	22	0
1	146	56	0	0	29.7	0.564	29	0
1	103	80	11	82	19.4	0.491	22	0
0	146	82	0	0	40.5	1.781	44	0
1	79	75	30	0	32	0.396	22	0

Table 3. First 6 samples of the sampled dataset

Firstly, do a density plot to see the shapes of each important column separated by the density of the outcome itself.



Figures 10 - 14. Series of Density Plots per Variables

As seen in the figures, As initial prediction, this study uses the actual values from the logistic regression model to gauge how accurate the empirical values from the permutation approach will be.

Predictor	Estimate	Std. Error	z-value	p-value	Significance
Intercept	-9.42584	2.13309	-4.419	9.92e-06	
Pregnancies	0.19166	0.09948	1.927	0.05404	.Significant
Glucose	0.04235	0.01127	3.757	0.00017	Significant
Blood Pressure	-0.01699	0.01702	-0.999	0.31792	Not Significant
Skin Thickness	0.02064	0.02133	0.968	0.33305	Not Significant
Insulin	-0.00036	0.00231	-0.157	0.87532	Not Significant
BMI	0.10046	0.04970	2.021	0.04325	Significant
Diabetes Pedigree Function	0.55688	0.85872	0.649	0.51666	Not Significant
Age	-0.00075	0.02894	-0.026	0.97937	Not Significant

Table 4. Summary of the whole simple regression model

Null Deviance: 127.371 (df = 98)

Residual Deviance: 85.479 (df = 90)

AIC: 103.48

As seen in the summary, Glucose had the highest significance in the odds of having diabetes, with a p-value of 0.0423498. The estimated coefficient of 0.043247 suggests that as a unit of glucose increases, the odds of having diabetes also increase, translating to higher probability of being diabetic.

BMI also resulted in moderate significance with a p-value of 0.043247. This indicates that as BMI increases by 1 unit, the odds also increase by an estimated coefficient of 0.1004562.

Pregnancies is borderline insignificant with a p-value of 0.054038, meaning there is some evidence of an effect, but not strong enough to meet the threshold. The rest of the features are insignificant and have no significant effect on the odds of having diabetes.

Since multiple logistic regression takes account of all columns on the p-values, we can check whether the p-value of each important feature will have a major change in their significance on the odds of having diabetes.

Firstly, Glucose

Predictor	Estimate	Std. Error	z-value	p-value	Significance
Intercept	-6.12553	1.26528	-4.841	1.29e-06	
Glucose	0.04398	0.00977	4.501	6.75e-06	Significant

Table 5. Summary of Intercept with Glucose

Null Deviance: 127.371 (df = 98)

Residual Deviance: 99.121 (df = 97)

AIC: 103.12

The summary resulted in Glucose being highly significant to the odds of having diabetes with a p-value of 6.75e-06. This justifies the previous model that also resulted in Glucose being highly significant. This indicates that despite the sharing of traits and coefficients in multiple logistic regression, Glucose remains highly significant in the odds of having diabetes.

Predictor	**Estimate**	**Std. Error**	**z-value**	**p-value**	**Significance**
Intercept	-4.28135	1.27143	-3.367	0.00076	
BMI	0.11031	0.03726	2.960	0.00307	**

Table 6. Summary of Intercept with BMI

Null Deviance:	127.37 (df = 98)
Residual Deviance:	116.38 (df = 97)
AIC:	120.38

Lastly, BMI is significant and has an effect on the odds of having diabetes with a p-value of 0.003073. This is slightly inclined with the multiple regression as BMI was mildly significant with a p-value of 0.043247. This indicates that although there was slight multicollinearity happening, BMI is still a significant factor in affecting the odds of having diabetes.

With this in mind, we do the Kolmogorov–Smirnov (KS) test using a permutation approach to find the empirical p-value to accept or reject the hypothesis.

Null Hypothesis: Pregnancies, Glucose, BloodPressure, Insulin and BMI does not affect the odds of having diabetes

Alternative Hypothesis: Pregnancies, Glucose, BloodPressure, Insulin and BMI has significant affect in the odds of having diabetes.

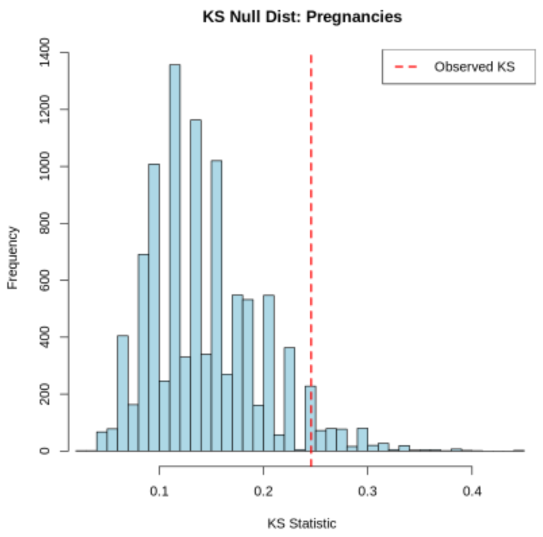
Variable	KS Statistic	p-value	Significance
Pregnancies	0.2457	0.0587	Borderline
Glucose	0.6122	0.0001	*** Significant
Blood Pressure	0.1837	0.2963	Not Significant
Insulin	0.3882	0.0013	** Significant
BMI	0.3100	0.0183	* Significant

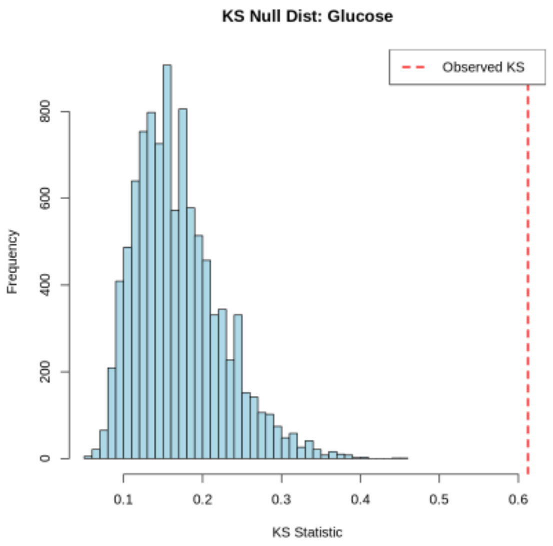
Table 7. Overall Summary

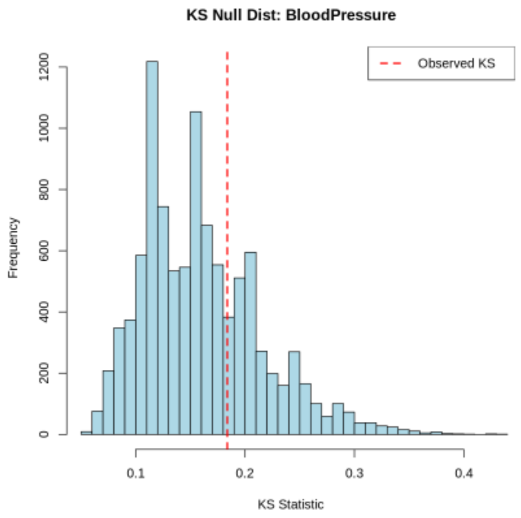
Glucose, Insulin and BMI reject the null hypothesis thus having significant effect in the odds of having diabetes with p-values of 0.00009999, 0.00069993, and 0.01768823 respectively. This indicates that these features have the highest effect on the odds with Glucose being the most impactful.

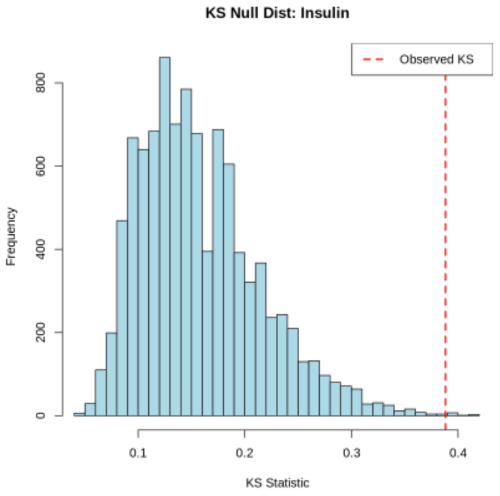
Pregnancies have borderline significance but not enough to reject the null hypothesis. In constrast, blood pressure had high p_value, thus showed no significant effect as well as reject the null hypothesis.

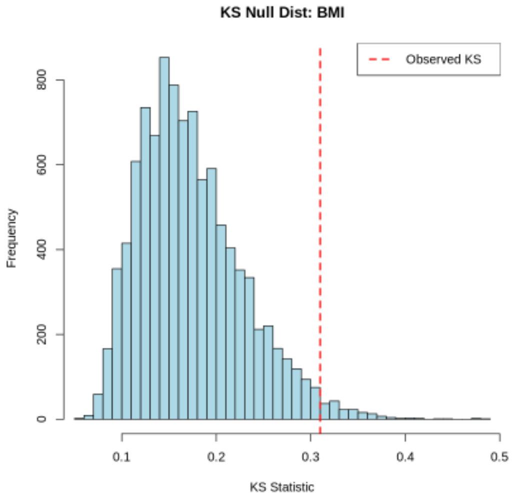
Overall, the findings from the permutation-based Kolmogorov–Smirnov tests align well with those from the individual simple logistic regression models. This makes sense as the permutation approach compares the distribution one feature at a time similar to a simple logistic regression model where it estimates the effect one feature at a time. The permutation approach successfully highlighted the same key predictors, supporting the legitimacy of its use in evaluating feature importance and effects.











Figures 15-19. KS Null Distribution Each Variables

Figure above is the collection of Null distribution of the key features. Firstly, the observed KS of Pregnancies lies within the tail of the distribution, indicating that while the observed difference between the distribution of individuals with and without diabetes is relatively large, it is still not sufficient to confidently reject the null hypothesis.

Glucose on the other hand has an observed KS at 0.6 there is at least 60% maximum difference between the cumulative distributions of Glucose with and without diabetes. This means that the Glucose level of individuals with and without diabetes are largely different, which supports the rejection of the null hypothesis.

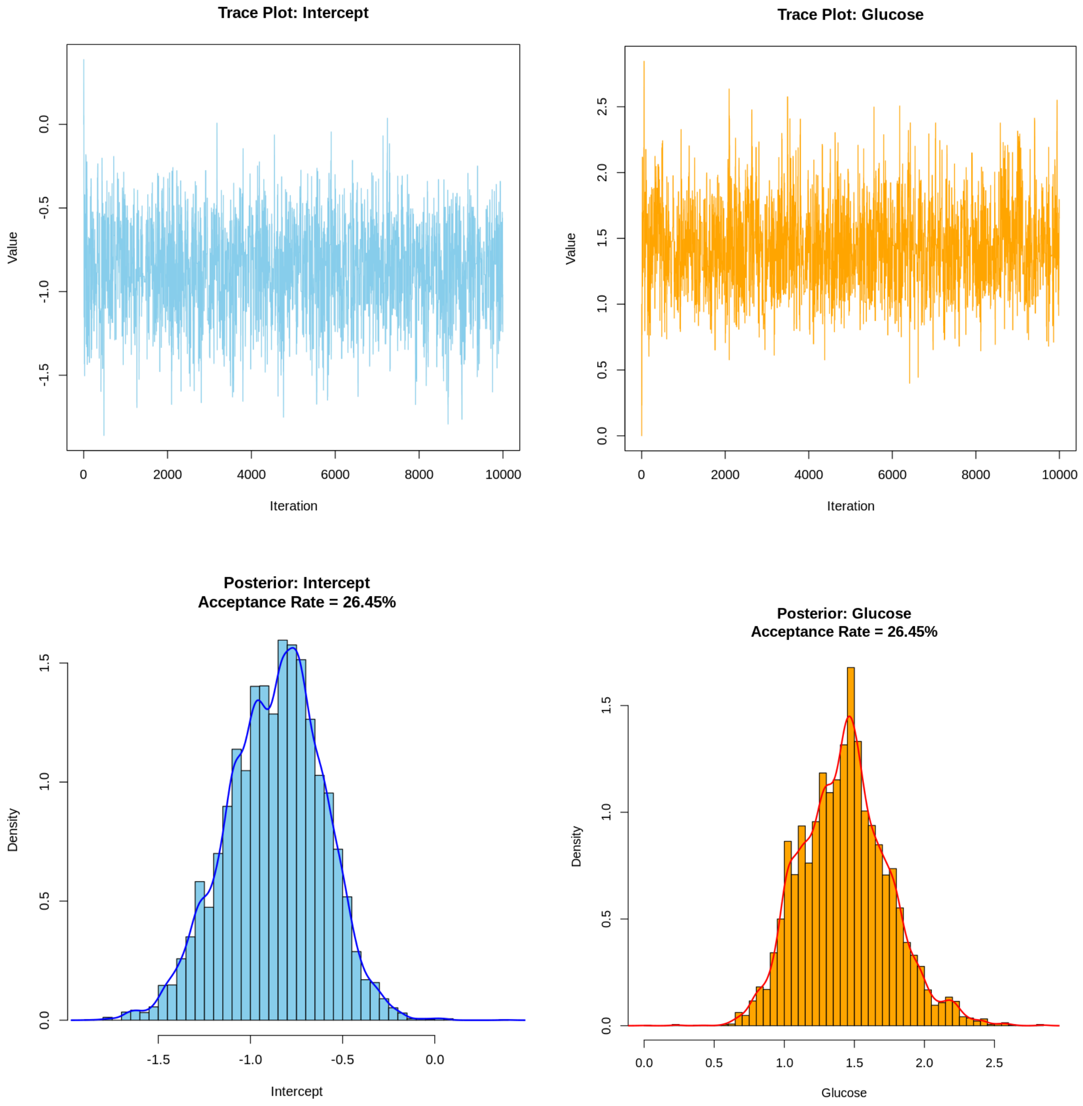
Blood pressure has the observed KS at approximately 0.18, indicating a very low difference between the cumulative distribution. This means that the blood pressure of individuals are not necessarily different.

Insulin has the observed KS at approximately 0.38, reflecting a moderate divergence of the distribution between individuals with and without diabetes. This suggests that individuals that are diabetic and non-diabetic have a relatively large difference on insulin levels.

BMI has the observed KS at approximately 0.31, also reflecting a moderate divergence of the distribution between individuals with and without diabetes. Similar to the Insulin, this suggests that the individuals that are diabetic and non-diabetic have a relatively large difference on BMI levels, although slightly less strongly than insulin.

Overall, the figures justify the empirical p-values given from the KS test using the permutation approach. Moreover, it is also justified with the initial multiple logistic regression made from

MCMC for Bayesian Inference



Figures 20 - 23. Trace Plot with Respective Posterior Density for Intercept and Glucose

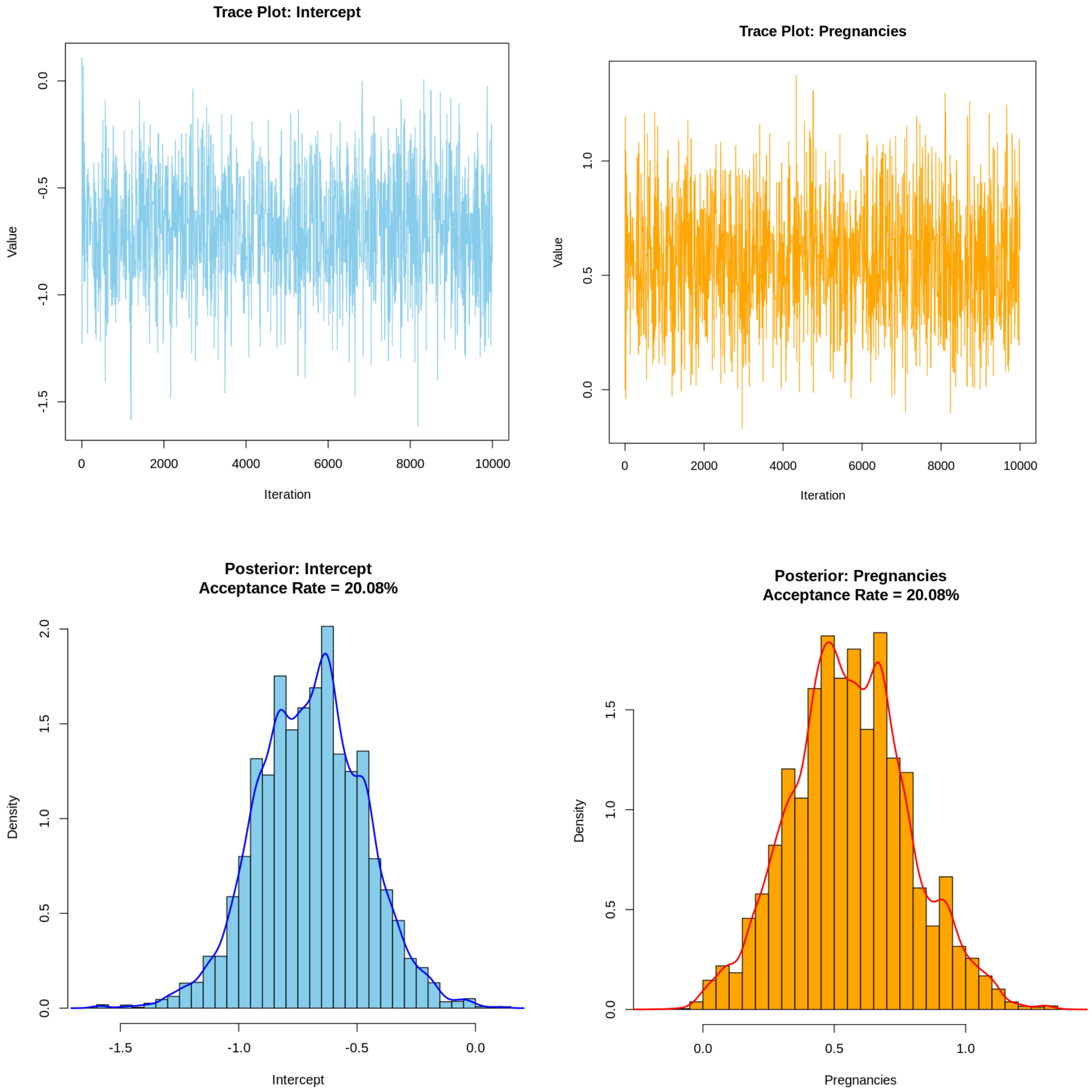
Figures above show the trace plots of the intercept and glucose as well as their posterior distributions. Firstly, the trace plot of the intercept fluctuates on the values near -1. Moreover, there is no visible drift or trend on the chain indicating that the samples are well-mixed and convergent.

Similarly, the trace plot of the glucose hovers around 1.5 with no visible stretches and trends which indicates a good exploration of the posterior. This means that both trace plots have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -1.2 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when glucose is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well.

On the other hand, the posterior distribution of the glucose's values centers around 1.5 indicating that the glucose have a high confidence positive relationship on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well.

Overall, the trace plots show a well-mixed and convergent posterior as well as having a roughly symmetric distribution with a good acceptance rate (Given that the MCMC used is Metropolis-Hastings).



Figures 24-27. Trace Plot with Respective Posterior Density for Intercept and Pregnancies

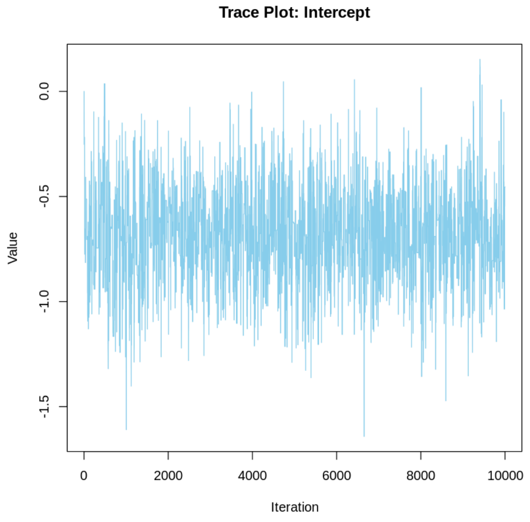
Figures above show the trace plots of the intercept and pregnancies as well as their posterior distributions. Firstly, the trace plot of the intercept fluctuates on the values near -0.7. Moreover, there is no visible drift or trend on the chain indicating that the sample are well-mixed and convergent.

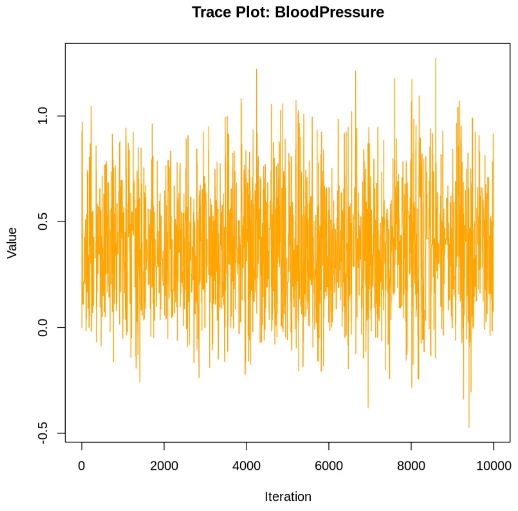
Similarly, the trace plot of the pregnancies hovers around 0.5 with no visible stretches and trends which indicates a good exploration of the posterior. This means that both trace plot have mixed well and are convergent.

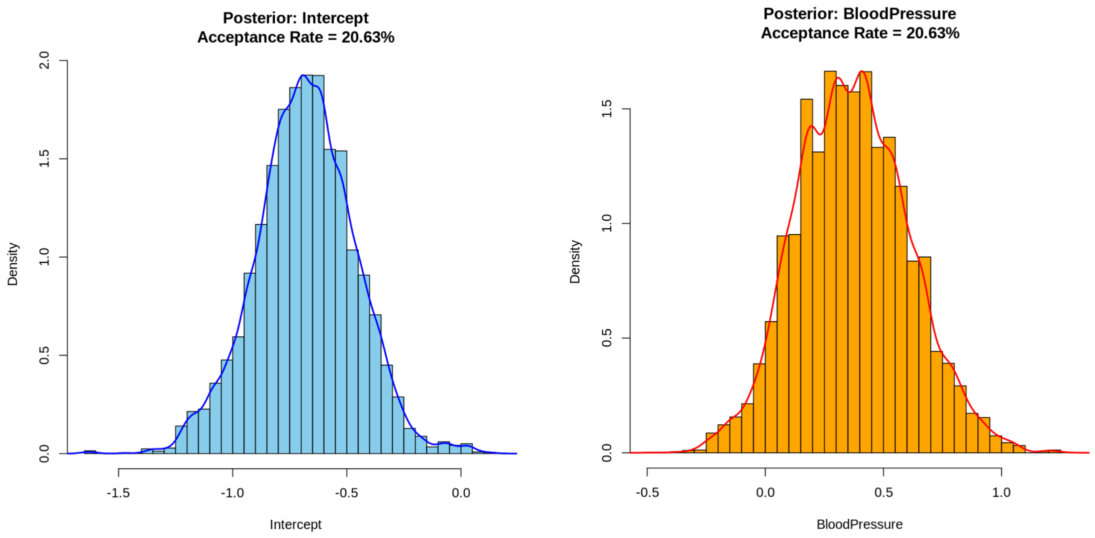
For the posterior distribution of the intercept, the values centers around -0.7 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when pregnancies is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well.

On the other hand, the posterior distribution of the pregnancy's values centers around 0.5 indicating that the Pregnancies have a high confidence positive relationship on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well. However, there is a shallow 2nd peak around 0.7 indicating a lumpy unimodal but generally still consistent.

Overall, the trace plots show a well-mixed and convergent posterior as well as having a roughly symmetric distribution with a good acceptance rate.


Trace Plot: Intercept


Trace Plot: BloodPressure



Figures 28 - 31. Trace Plot with Respective Posterior Density for Intercept and Blood Pressure

Figures above show the trace plots of the intercept and BloodPressure as well as their posterior distributions. Firstly, the trace plot of the intercept fluctuates on the values near -0.7. Moreover, there is no visible drift or trend on the chain indicating that the samples are well-mixed and convergent.

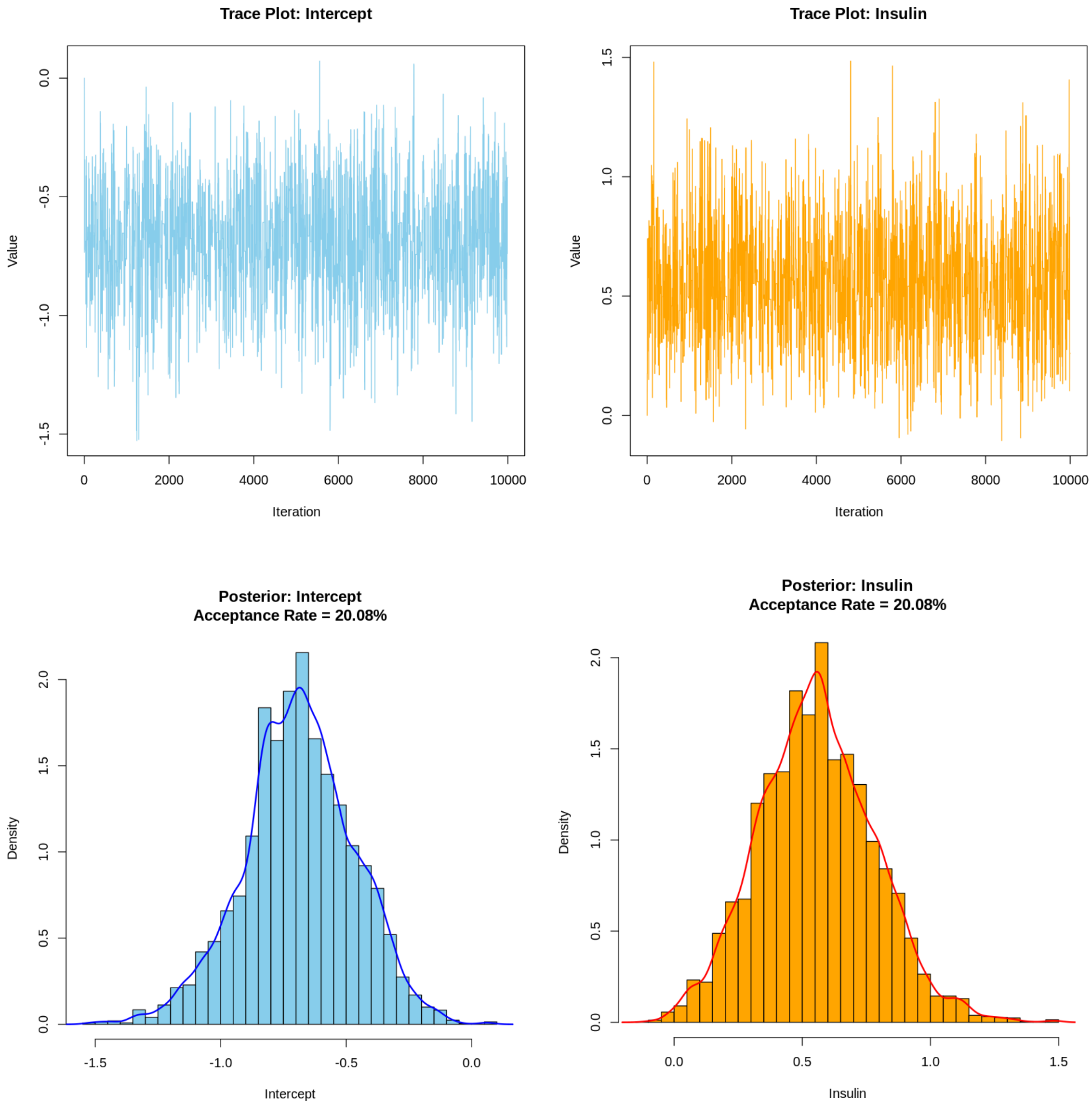
Similarly, the trace plot of the BloodPressure hovers around 0.3 with no visible stretches and trends which indicates a good exploration of the posterior. This means that both trace plot have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -0.7 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when BloodPressure is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well. However, there is a shallow 2nd peak as well around 0.6 indicating a lumpy unimodal but generally still consistent.

On the other hand, the posterior distribution of the BloodPressure's values centers around 0.3 indicating that the BloodPressure has a positive relationship on the odds of having diabetes.

However, it may still be uncertain or not entirely confident since the empirical p-value shows the BloodPressure as insignificant. Therefore, there is still not enough to conclude significance to whether it has an effect on diabetes.

Overall, the trace plots show a well-mixed and convergent posterior as well as having a roughly symmetric distribution with a good acceptance rate.



Figures 32 - 35. Trace Plot with Respective Posterior Density for Intercept and Insulin

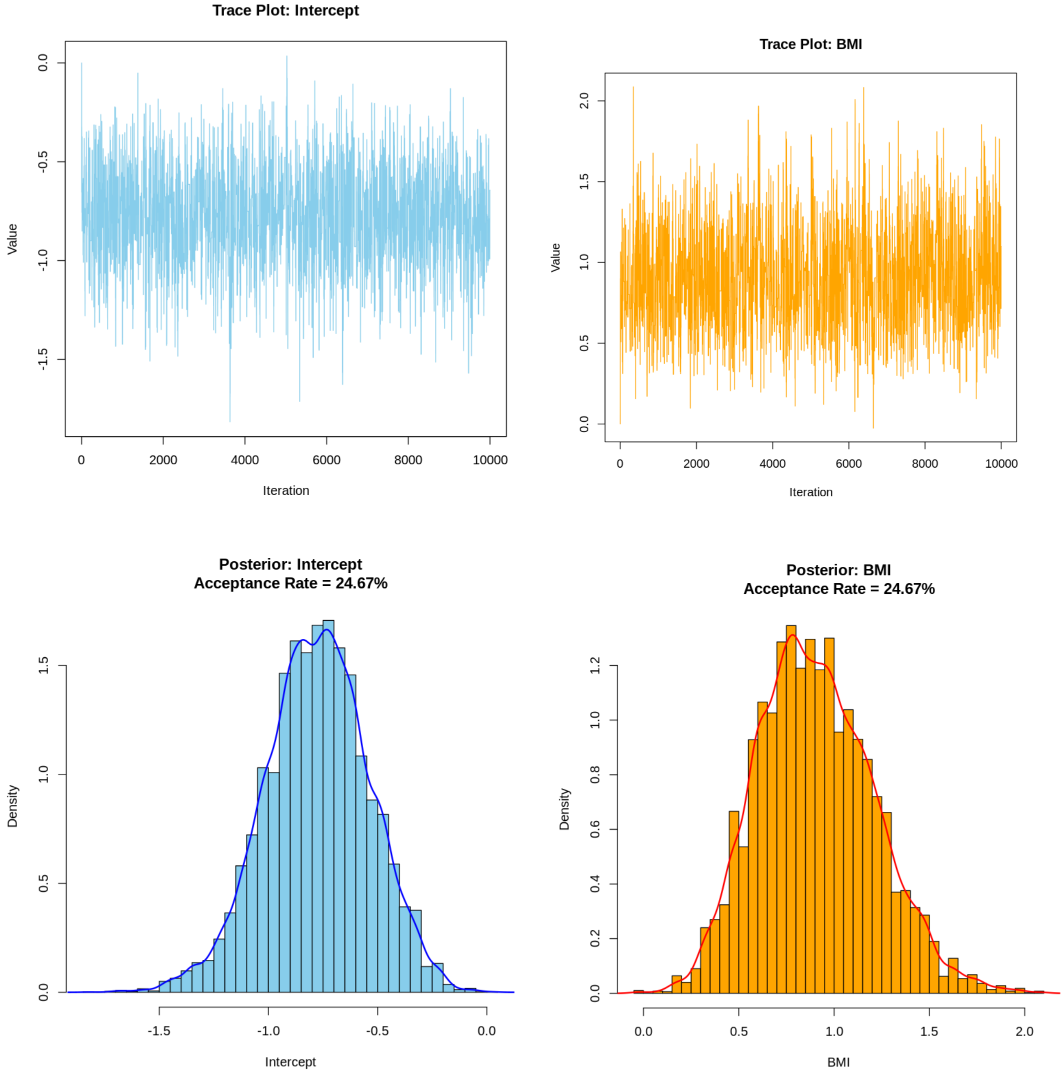
Figures above show the trace plots of the intercept and Insulin as well as their posterior distributions. Firstly, the trace plot of the intercept fluctuates on the values near -0.6. Moreover, there is no visible drift or trend on the chain indicating that the samples are well-mixed and convergent.

Similarly, the trace plot of the Insulin hovers around 0.5 with no visible stretches and trends which indicates a good exploration of the posterior. This means that both trace plot have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -0.6 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when Insulin is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well.

On the other hand, the posterior distribution of the Insulin's values centers around 0.5 indicating that the Insulin has a high confidence positive relationship on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well.

Overall, the trace plots show a well-mixed and convergent posterior as well as having a roughly symmetric distribution with a good acceptance rate.



Figures 36 - 39. Trace Plot with Respective Posterior Density for Intercept and BMI

Figures above show the trace plots of the intercept and BMI as well as their posterior distributions. Firstly, the trace plot of the intercept fluctuates on the values near -0.7. Moreover, there is no visible drift or trend on the chain indicating that the samples are well-mixed and convergent.

Similarly, the trace plot of the BMI hovers around 0.8 with no visible stretches and trends which indicates a good exploration of the posterior. This means that both trace plots have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -0.7 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when BMI is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well. However, there is a tiny 2nd peak indicating a slight lump on the distribution, but generally still consistent.

On the other hand, the posterior distribution of the BMI's values centers around 0.8 indicating that the BMI have a high confidence positive relationship on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well.

Overall, the trace plots show a well-mixed and convergent posterior as well as having a roughly symmetric distribution with a good acceptance rate.

With that, we can also compare the result of the Metropolis-Hasting algorithm to a different algorithm, specifically the Hamiltonian Monte Carlo where it uses gradient-based proposals instead of random walk from the MC algorithm. This means that it leverages the shape and slope to make an informed movement across the entire posterior. With this, it makes the exploration better and faster compared to the MH algorithm.

Bayesian Logistic Regression Output

Family: Bernoulli

Link Function: Logit

Formula: Outcome ~ scale(variable)

Data Used: df_sampled (n = 99)

Sampling Method: NUTS (No-U-Turn Sampler)

Chains: 4

Iterations per chain: 2000 (with 1000 warmup)

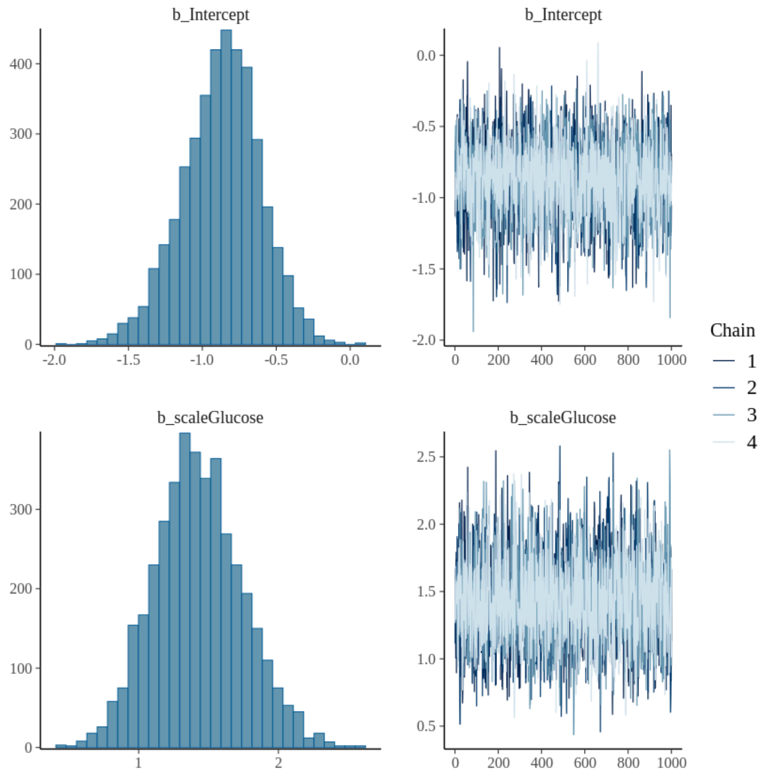
Total post-warmup draws: 4000

Parameter	Estimate	Est. Error	95% CI (Lower)	95% CI (Upper)	R̂	Bulk_ESS	Tail_ESS
Intercept	-0.87	0.27	-1.42	-0.37	1	2802	2314
scale(Glucose)	1.43	0.32	0.84	2.09	1	2982	2619

Table 8. Bayesian Logistic Regression Coefficients (Glucose)

Notes:

- R̂ (R-hat) near 1.00 indicates convergence across chains.
- Bulk_ESS and Tail_ESS are effective sample size metrics; higher is better.



Figures 40 - 43. Trace Plot and Posterior Distribution of Intercept and Glucose

Figures above show the trace plots of the intercept and Glucose as well as their posterior distributions. Firstly, the summary shows that the scaleGlucose has a positive relationship indicating that for every 1 sd increase in Glucose, the log odds increase by 1.44 of having diabetes. Moreover, the intercept has a negative relationship indicating that when Glucose is 0, the log-odds of having diabetes is -0.87.

The trace plot of the intercept fluctuates on the values near -0.9 on all 4 chains. Moreover, there is no visible drift or trend on the chain indicating that the all 4 chains are well-mixed and convergent.

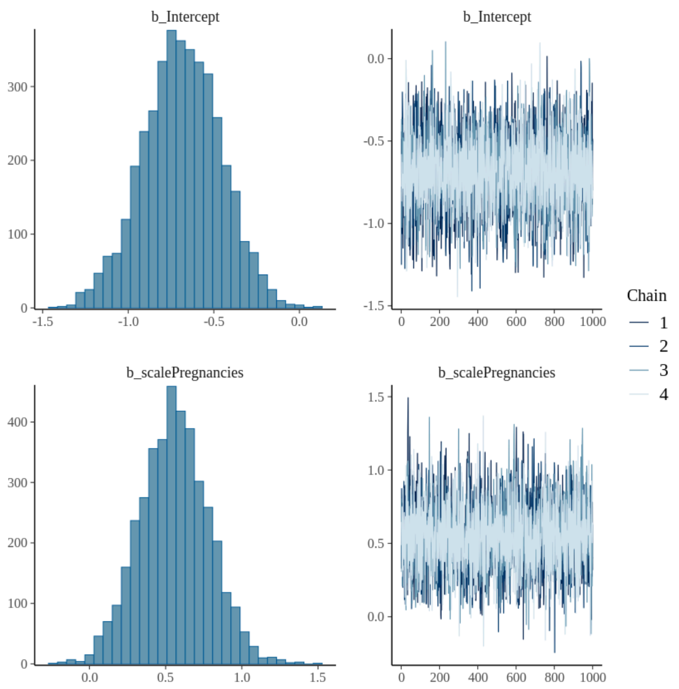
Similarly, the trace plot of the Glucose hovers around 1.5 with no visible stretches and trends on all 4 chains which indicates a good exploration of the posterior. This means that both trace plot have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -0.8 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when Glucose is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well.

On the other hand, the posterior distribution of the Glucose's values centers around 1.5 indicating that the Glucose has a high confidence positive relationship on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well.

Parameter	Estimate	Est. Error	95% CI (Lower)	95% CI (Upper)	$\hat{R}$	Bulk_ESS	Tail_ESS
Intercept	-0.69	0.23	-1.15	-0.25	1	3064	2319
scale(Pregnancies)	0.56	0.23	0.11	1.01	1	3286	2613

Table 9. Bayesian Logistic Regression Coefficients (Pregnancies)



Figures 44 - 47. Trace Plot and Posterior Distribution of Intercept and Glucose

Figures above show the trace plots of the intercept and Pregnancies as well as their posterior distributions. Firstly, the summary shows that the scalePregnancies have a positive relationship indicating that for every 1 sd increase in Pregnancies, the log odds increase by 0.55 of having diabetes. Moreover, the intercept has a negative relationship indicating that when Pregnancies is 0, the log-odds of having diabetes is -0.69.

The trace plot of the intercept fluctuates on the values near -0.6 on all 4 chains. Moreover, there is no visible drift or trend on the chain indicating that the all 4 chains are well-mixed and convergent.

Similarly, the trace plot of the Pregnancies hovers around 0.6 with no visible stretches and trends on all 4 chains which indicates a good exploration of the posterior. This means that both trace plots have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -0.6 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when Pregnancies is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well.

On the other hand, the posterior distribution of the Pregnancy's values centers around 0.6 indicating that the Pregnancies have a high confidence positive relationship on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well.

Parameter	Estimate	Est. Error	95% CI (Lower)	95% CI (Upper)	$\hat{R}$	Bulk_ESS	Tail_ESS
Intercept	-0.68	0.22	-1.12	-0.24	1	3247	2509
scale(Blood Pressure)	0.38	0.24	-0.07	0.87	1	3237	2565

Table 10. Bayesian Logistic Regression Coefficients (Blood Pressure)

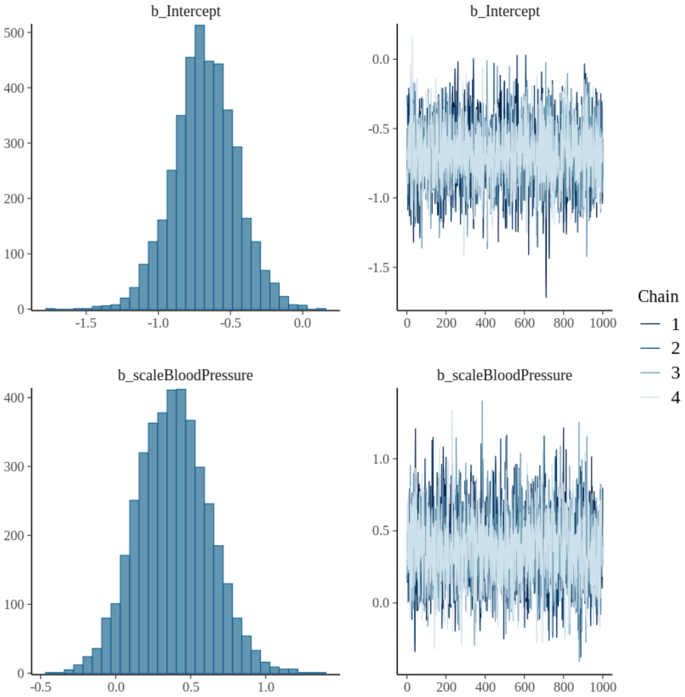


Figure 48 - 51. Trace Plot and Posterior Distribution of Intercept and Blood Pressure

Figures above show the trace plots of the intercept and BloodPressure as well as their posterior distributions. Firstly, the summary shows that the scaleBloodPressure have a positive relationship indicating that for every 1 sd increase in Pregnancies, the log odds increase by 0.39 of having diabetes. However, since the empirical p-value result from the permutation approach is insignificant, it is not confident that the BloodPressure feature necessarily has an effect. Moreover, the intercept has a negative relationship indicating that when BloodPressure is 0, the log-odds of having diabetes is -0.69.

The trace plot of the intercept fluctuates on the values near -0.7 on all 4 chains. Moreover, there is no visible drift or trend on the chain indicating that the all 4 chains are well-mixed and convergent.

Similarly, the trace plot of the BloodPressure hovers around 0.3 with no visible stretches and trends on all 4 chains which indicates a good exploration of the posterior. This means that both trace plot have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -0.7 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when BloodPressure is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well.

On the other hand, the posterior distribution of the BloodPressure's values centers around 0.3 indicating that the BloodPressure have a positive relationship on the odds of having diabetes. With that, due to the result of the empirical p-value, it is not with confidence that the BloodPressure have an effect on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well.

Parameter	Estimate	Est. Error	95% CI (Lower)	95% CI (Upper)	R̂	Bulk_ESS	Tail_ESS
Intercept	-0.68	0.23	-1.12	-0.25	1	3300	2335
scale(Insulin)	0.56	0.23	0.12	1.03	1	3441	2378

Table 11. Bayesian Logistic Regression Coefficients (Insulin)

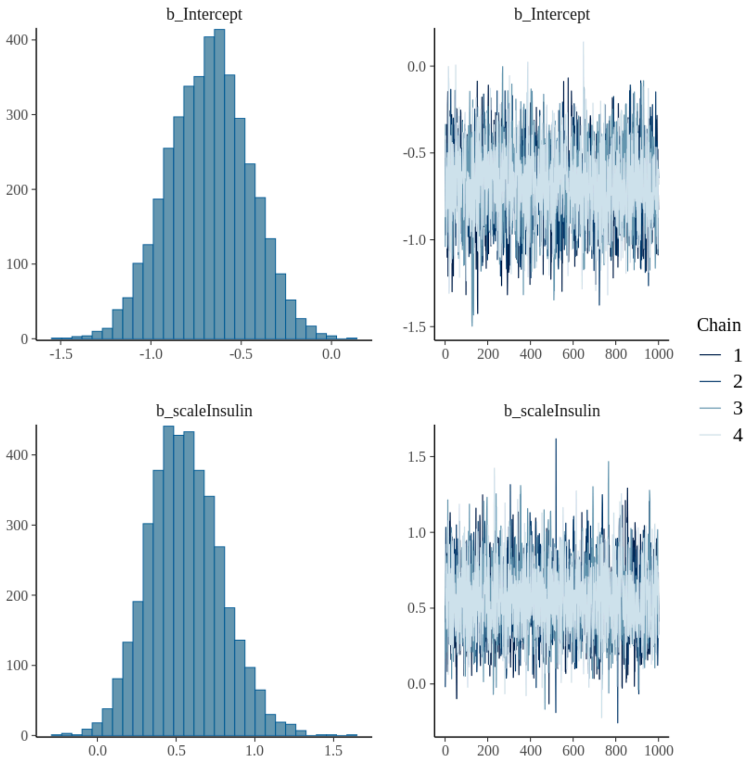


Figure 52 - 55. Trace Plot and Posterior Distribution of Intercept and Insulin

Figures above show the trace plots of the intercept and Insulin as well as their posterior distributions. Firstly, the summary shows that the scaleInsulin has a positive relationship indicating that for every 1 sd increase in Insulin, the log odds increase by 0.56 of having diabetes. Moreover, the intercept has a negative relationship indicating that when Insulin is 0, the log-odds of having diabetes is -0.68.

The trace plot of the intercept fluctuates on the values near -0.6 on all 4 chains. Moreover, there is no visible drift or trend on the chain indicating that the all 4 chains are well-mixed and convergent.

Similarly, the trace plot of the Insulin hovers around 0.6 with no visible stretches and trends on all 4 chains which indicates a good exploration of the posterior. This means that both trace plot have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -0.6 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when Insulin is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well.

On the other hand, the posterior distribution of the Insulin's values centers around 0.6 indicating that the Insulin has a high confidence positive relationship on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well.

BMI

Parameter	Estimate	Est. Error	95% CI (Lower)	95% CI (Upper)	$\hat{R}$	Bulk_ESS	Tail_ESS
Intercept	-0.77	0.24	-1.26	-0.33	1	2689	2595
scale(BMI)	0.89	0.29	0.33	1.49	1	2889	2598

Table 11. Bayesian Logistic Regression Coefficients (BMI)

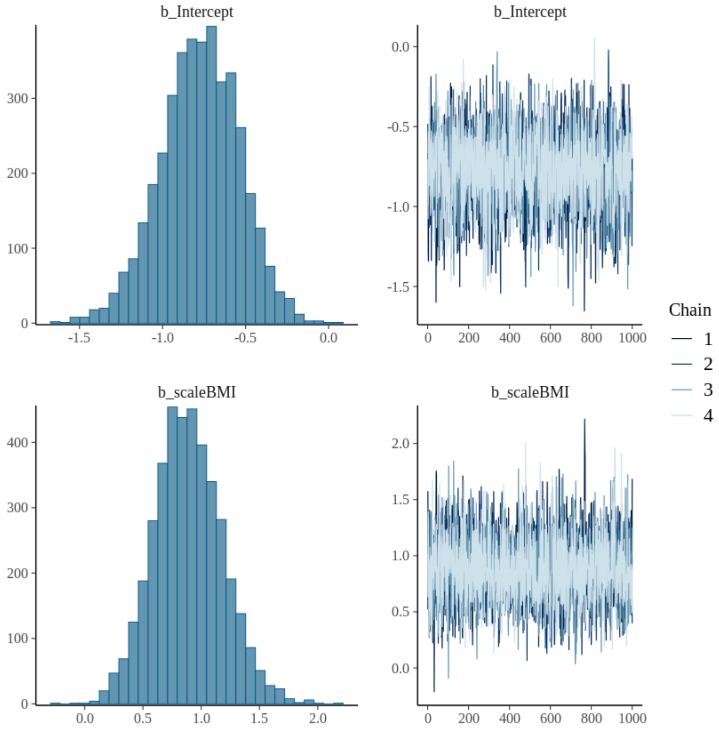


Figure 56 - 59. Trace Plot and Posterior Distribution of Intercept and BMI

Figures above show the trace plots of the intercept and BMI as well as their posterior distributions. Firstly, the summary shows that the scaleBMI have a positive relationship indicating that for every 1 sd increase in BMI, the log odds increase by 0.90 of having diabetes. Moreover, the intercept has a negative relationship indicating that when BMI is 0, the log-odds of having diabetes is -0.78.

The trace plot of the intercept fluctuates on the values near -0.6 on all 4 chains. Moreover, there is no visible drift or trend on the chain indicating that the all 4 chains are well-mixed and convergent.

Similarly, the trace plot of the BMI hovers around 0.9 with no visible stretches and trends on all 4 chains which indicates a good exploration of the posterior. This means that both trace plots have mixed well and are convergent.

For the posterior distribution of the intercept, the values centers around -0.8 indicating that the intercept has a high confidence negative relationship on the odds of having diabetes when BMI is 0. Moreover, the distribution is unimodal and roughly symmetric with the density curve fitting well.

On the other hand, the posterior distribution of the BMI's values centers around 0.9 indicating that the Insulin has a high confidence positive relationship on the odds of having diabetes. Similar to the intercept, the distribution is unimodal and roughly symmetric with the density curve fitting well.

Overall, the HMC is better at detecting noise thus having a much clearer figure on all the posterior distributions, while the MH had a bit of a lumpy structure on some features. Moreover, HMC uses less computational value whilst keeping it truthful on all 4 chains. This is because the model only used n=2000, looping into 4 chains equating to 8000 loops, however 4000 of those are discarded as warm up. On the other hand, MH used 10000 iterations to properly create 1 chain. The informed movement of the HMC makes it much more efficient, faster and more accurate than the MH algorithm.

Bootstrap and Jackknife

Although other parts of the study have explored inference and resampling methods in the early process to filter out and identify which features are most impactful, the bootstrap and jackknife of this study is to focus on the highest p-value that drives the susceptibility of Diabetes. This study aims to examine glucose (p-value = 6.75e-06) and its stability and bias of its estimated mean and regression coefficient.

Although this is the case, task 4: permutation tests reveal that there are other predictors that are significant on the 0.05 p-value threshold. In addition to the glucose, two more predictors, namely BMI (p-value = 0.003073) and Insulin (p-value = 0.017341), will also be subject to the resampling method as an additional predictor to the glucose. Permutation tests identified glucose as highly significant, hence this became the main focus for bias/variance estimation via bootstrap and jackknife.

Term	Estimate	Std. Error	z-value	p-value	Significance
Intercept	-5.35008	0.42083	-12.71	< 2e-16	Significant
Glucose	0.03787	0.00325	11.65	< 2e-16	Significant

Table 12. Bootstrapping Logistic Regression Coefficients (Glucose)

Null deviance: 993.48 on 767 degrees of freedom
Residual deviance: 808.72 on 766 degrees of freedom
AIC: 812.72

Logistic Regression Analysis of Glucose on Diabetes

Using the independent variable of glucose on the given dataset, a logistic regression analysis was conducted to examine the relationship between glucose levels and the likelihood of having diabetes among pregnant women. The model was fitted using a binary logistic regression approach, where the dependent variable was Outcome.

The result of the analysis showed that glucose level is a statistically significant predictor of diabetes status. The estimated coefficient for glucose was 0.03787 with a standard error of 0.00325. This coefficient was highly significant, backed by the p-value less than 0.0001, indicating a very strong relationship between glucose levels and the probability of having diabetes. Based on the coefficient, it translates that pregnant ladies increase the chance of having diabetes by around 3.9% per unit in glucose, after converting the log-odds to odds ratios using the exponential function. Since this model only includes a single predictor, which in this case, glucose, the AIC for the model shows an evaluation of 812.72, suggesting a reasonable balance between goodness of fit and model. In totality, the findings of the model suggest a strong and statistically significant association and the exact possibility of having diabetes.

The dataset could further be explained through a series of visualization.

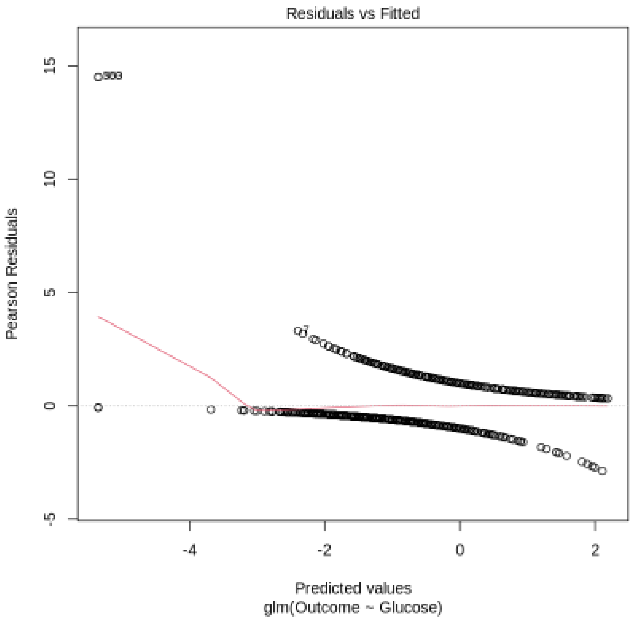


Figure 60. Residuals vs Fitted

To evaluate the logistic regression model based on glucose levels, this study examines the four standard diagnostic plots generated.

Figure 60 tackles the Residuals vs. Fitted plot. The data above shows the spread of deviance residuals against the predicted value. In this case, the residuals generally center around zero but there are noticeable trends and some curvature, especially at the extremes of the predicted values. This means that there is potential non-linearity or a variable bias, indicating that glucose alone may not be able to capture the true complexity of the diabetes.

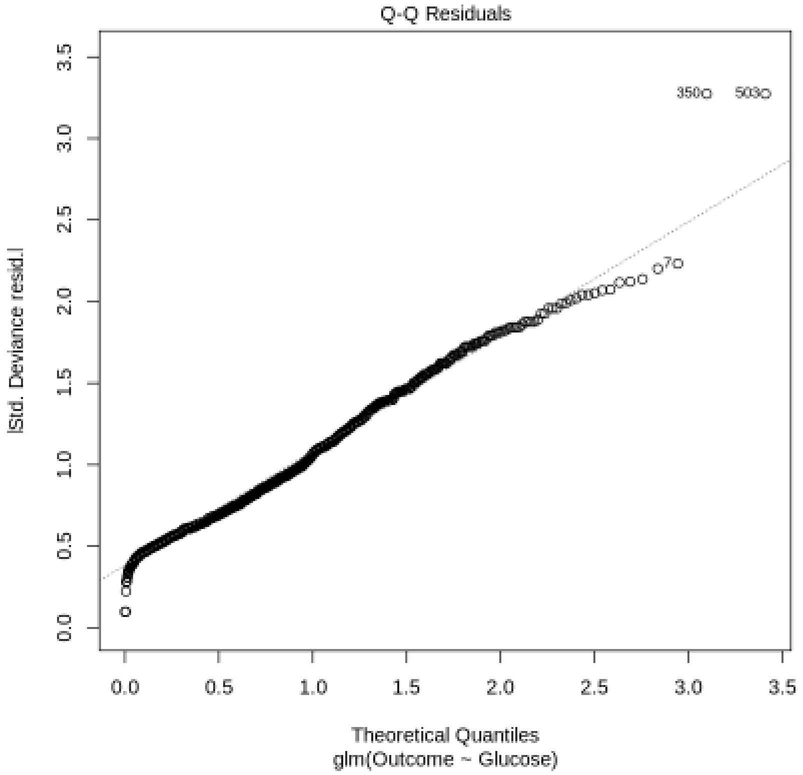


Figure 61. Q-Q Residuals

Figure 61 shows the Normal q-q plot. The data above shows a pattern that residuals are not perfectly distributed among the range, and most of the values are secluded around 0 to 2. Although this is the case, this is somewhat expected in logistic regression due to the binary nature of the outcome.

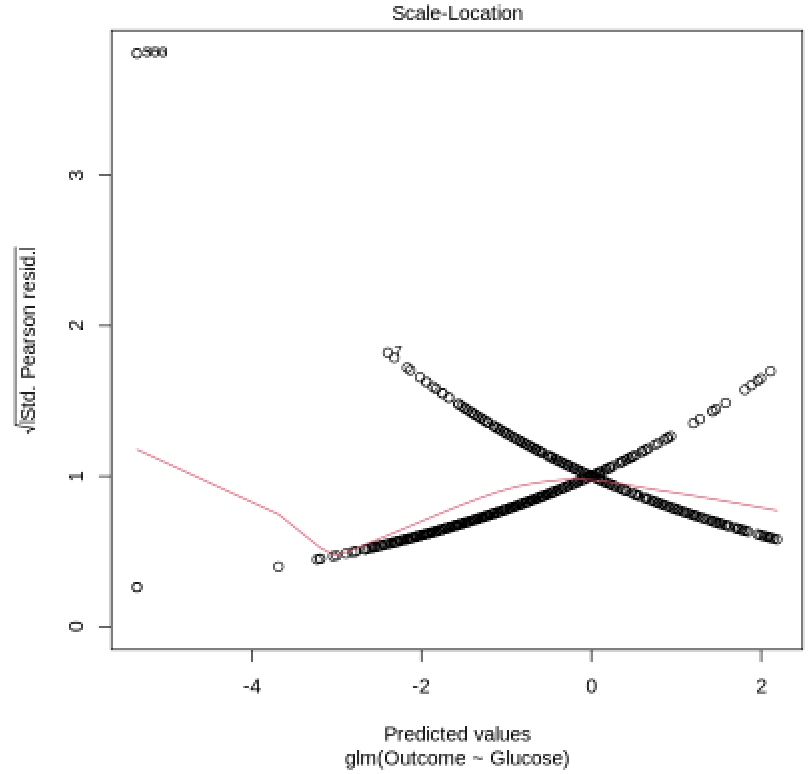


Figure 62. Scale-Location

Figure 62 describes the Scale-Location plot. Based on the figure, it appears that there is an "X" shaped pattern, which could indicate heteroscedasticity or unequal variance of residuals across predicted values. This is the cause because the model fit may be improved by more predictors, since residual spread increases at both low and high ends of predicted probabilities.

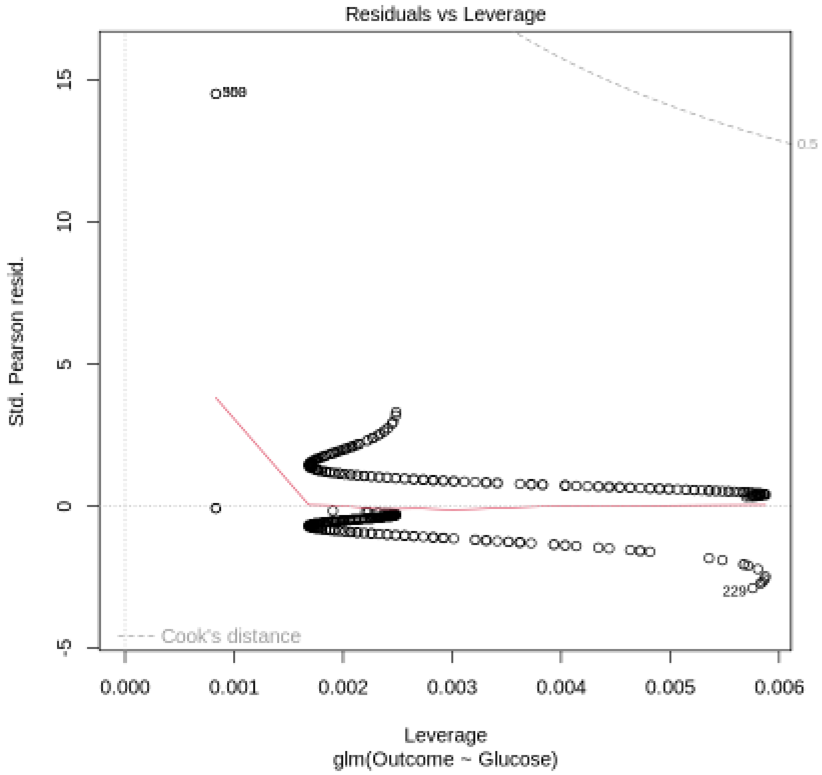


Figure 63. Residuals vs Leverage

Lastly, figure 63 shows the Residuals vs Leverage plot. Based on the graph, although most observations cluster near the center with low leverage, a small number of points are shown as outliers which provides a small influence on the model estimates.

Bootstrap Resampling

Following the diagnostic assessment of the logistic regression model, it was evident that there may be indicators of potential variance.

To further assess the stability and reliability of the estimated regression coefficient, since the data shows a limited amount, this study applies the bootstrap resampling method using 10,000 resamples.

Mean Coefficient: 0.03811

Bias Estimate: 0.000237

Variance Estimate: 1.4e-05

Using bootstrap resampling with glucose as the only predictor, this study focused on the mean coefficient estimate, bias, and variance of the logistic regression coefficient. The results indicated a mean coefficient of 0.03811, which is closely aligned with the original coefficient from the logistic regression model (approximately 0.03787). This suggests that the coefficient estimate is consistent across multiple resamples of the dataset.

In addition, the bias estimate shows to be 0.000237, indicating a very small value for bias, reinforcing the reliability of the coefficient.

Lastly, the variance estimate was 1.4e-04, reflecting a relatively low variance in the coefficient across bootstrap samples. This suggests that the regression coefficient remains stable after 10,000 resamples, strengthening the conclusion of the reliability of the glucose.

Although the value shows a reliable estimate, further analysis could expand to explore a higher chance of a predictor. Based on the permutation testing from the previous task, apart from the glucose, there are also significant variables that could help the model predict with better accuracy.

Bootstrap (w BMI)

Original Coefficient (w BMI): 0.03516896
Bootstrap Mean Coefficient (w BMI): 0.03536521
Bootstrap Bias Estimate (w BMI): 0.0001962528
Bootstrap Variance Estimate (w BMI): 1.364436e-05

A bootstrap resampling procedure with 10,000 iterations was applied to the logistic regression model incorporating both Glucose and Body Mass Index (BMI) as predictors of diabetes risk.

The results show that the bootstrap mean coefficient is around 0.035. In addition, the bias is, similar to the bootstrap only, shows very minimal bias at 0.000196. There is also a low spread, evident to the variance estimate.

Although changes are extremely minimal, the inclusion of BMI alongside glucose in the predictive model does not introduce instability to the glucose coefficient. In addition, it also slightly adjusts the magnitude of the coefficient.

Bootstrap (w Insulin)

Original Coefficient (w Insulin): 0.03888507
Bootstrap Mean Coefficient (w Insulin): 0.03916092
Bootstrap Bias Estimate (w): 0.0002758537
Bootstrap Variance Estimate (w Insulin): 1.532524e-05

In addition to BMI, another key biological factor often associated with diabetes risk is Insulin level. To explore its impact on the model's stability, a similar bootstrap procedure with 10,000 resamples was conducted for a logistic regression model.

With only a slightly higher evaluation than the two, the results indicate that the original coefficient for Glucose in this expanded model was 0.03889, while the bootstrap mean coefficient was 0.03916. In addition, the bias estimate found to be at 0.00028, which although a slight increase, is still a very acceptable bias rate. The variance estimate was 1.53×10^{-5}, reflecting low variability in the coefficient values across the bootstrap samples.

In summary, the addition of Insulin alongside Glucose results in a slightly higher average coefficient for glucose but maintains low bias and variance, supporting the reliability of the model. This suggests that both glucose and insulin are important contributors to the prediction of diabetes risk, and their combined inclusion enhances the robustness of the predictive model without compromising the stability of the estimates.

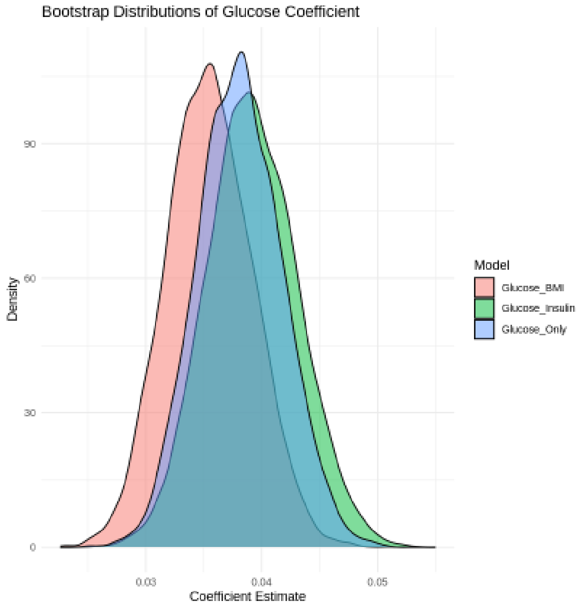


Figure 64. Bootstrap Coefficient Distribution Visualization

To compare the effect of adding different predictors apart from glucose, this plot illustrates the distribution of glucose coefficient estimates across 10,000 bootstrap resamples for three models: Glucose Only, Glucose + BMI, and Glucose + Insulin. Looking at the plot, it is evident that all three are relatively centered around the same coefficient, which affirms the positive association of glucose with the risk of diabetes for pregnant women. The Glucose + BMI shows a lower average glucose coefficient but is tighter on the distribution, meaning that it probably has the lowest variance and best stabilization. On the other hand, the glucose only shows a wider spread than the latter, which aligns with the earlier numerical findings where variance was larger in the univariate model. Lastly, the Glucose + Insulin model shows a slight shift to the right, indicating a higher average coefficient.

Overall, the visual comparison supports the conclusion that the inclusion of BMI reduces variance in the glucose coefficient, promoting stability, while inclusion of Insulin slightly increases the average effect size of glucose.

Compare with Theoretical Samples

To further evaluate the stability and reliability of the glucose coefficient under different model specifications, a comparative analysis was conducted between the original (theoretical) coefficient estimates derived from logistic regression and the corresponding bootstrap standard errors and confidence intervals.

Original (theoretical) Coefficient Estimates

Glucose only: 0.03787304

Glucose + BMI: 0.03516896

Glucose + Insulin: 0.03888507

The theoretical estimates obtained from the logistic regression models reveal consistent positives across all three. In the Glucose Only model, the coefficient was 0.03787, establishing a strong baseline association between glucose levels and diabetes risk. With the addition of BMI, the coefficient slightly decreased to 0.03517, indicating a modest reduction in the effect of glucose. Conversely, when Insulin was included as a covariate, the glucose coefficient slightly increased to 0.03889, suggesting that insulin adjustments slightly elevate the estimated effect of glucose on diabetes risk.

Bootstrap Standard Errors:

Glucose only: 0.003646946

Glucose + BMI: 0.003682998

Glucose + Insulin: 0.003923802

On the standard error, the glucose only model shows an evaluation of 0.00365, indicating moderate variability. Although still lower, the inclusion of BMI slightly increased to 0.00368, reflecting minimal additional variability introduced by adjusting for BMI. With the insulin, the standard error rose to about 0.00392.

This pattern implies that the glucose effect is more stable when controlling for BMI and more variable when insulin is considered, possibly due to physiological overlaps or interdependencies between glucose and insulin levels.

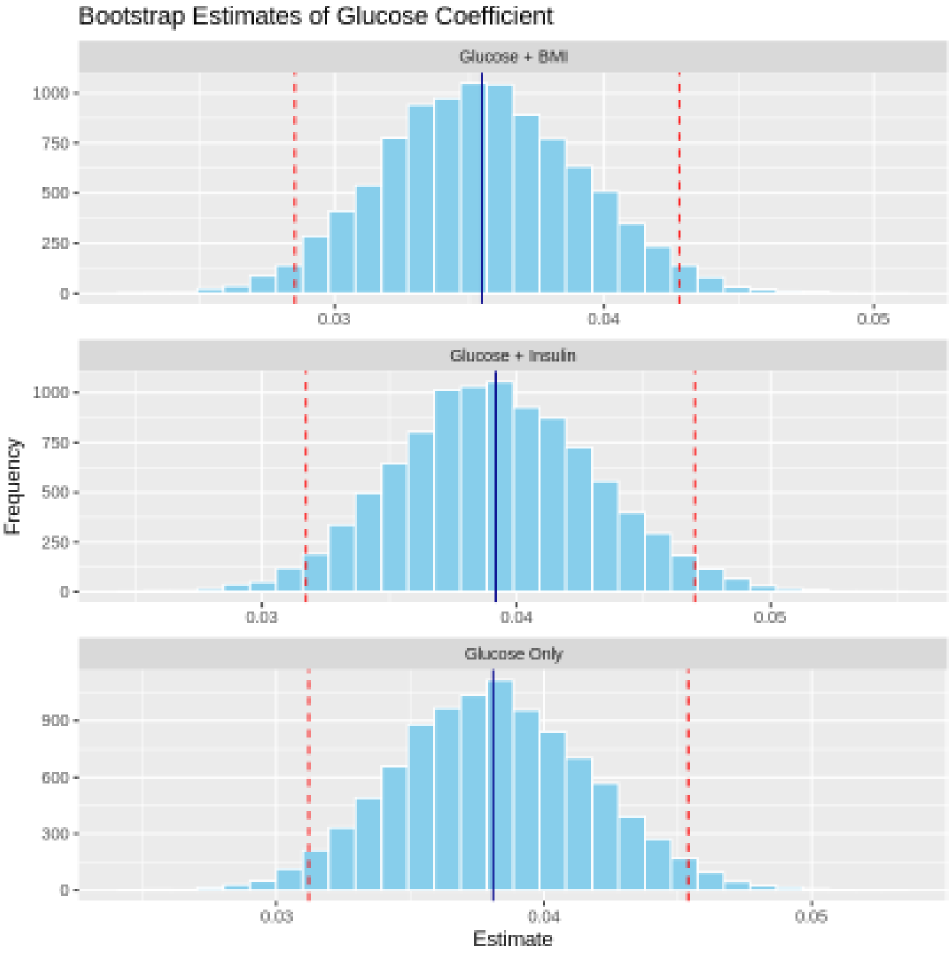
Bootstrap 95% Confidence Intervals (Percentile Method):

Glucose only: [0.03121598 - 0.04537643]

Glucose + BMI: [0.02852419 - 0.04278519]

Glucose + Insulin: [0.03172055 - 0.04702363]

For the Confidence Interval, it compares the three intervals at a 95% confidence. Looking at the ranges, it is clear that all three are almost similar in interval. Looking at the values, this shows that including BMI reduces both the point estimate and the range while stabilizing the glucose effect. On the other hand, including Insulin increases the point estimate and widens the interval, suggesting greater uncertainty but potentially a stronger conditional association between glucose and diabetes risk.



Figures 65-67. Bootstrap Estimate of Glucose Coefficient

To easily visualize the earlier theoretical parts mentioned, this shows a visualization among the three models. It is clear that the inclusion of additional predictors altered the magnitude, as well as the stability for the glucose coefficient. These findings reinforce the importance of model specification in risk prediction and suggest that while glucose remains a consistently strong predictor, its effect is influenced by the presence of other key physiological markers.

Jackknife

To complement the bootstrap analysis and further examine the stability of the glucose coefficient, a Jackknife resampling technique was employed. This approach provides an alternative way of estimating bias and variance, particularly useful for smaller datasets where evaluating the influence of each observation is critical.

Jackknife Glucose Statistic:

Original Coefficient: 0.03787304

Jackknife Mean Coef: 0.03787334

Jackknife Bias Estimate: 0.0002308984

Jackknife Variance: 1.342065e-05

Using Jackknife resampling, the glucose coefficient remained highly stable. The mean coefficient was 0.03787, nearly identical to the original estimate (0.03787), indicating no strong influence from any single data point. The bias was minimal (0.00023) and the variance low (1.34×10^{-5}), confirming the stability and reliability of glucose as a predictor of diabetes risk in this simple model. These results are consistent with the bootstrap findings, reinforcing the robustness of the coefficient.

Jackknife Glucose w/ BMI Statistic:

Jackknife Mean Coef (Glucose w BMI): 0.03516934

Jackknife Bias Estimate (Glucose w BMI): 0.0002922572

Jackknife Variance Estimate (Glucose w BMI): 1.367933e-05

Next, in the model including glucose and BMI, the jackknife coefficient is almosy similar with the original estimate of 0.031517, which indicates the strongest stability out of all. The bias remained small at 0.00029, and the variance was 1.37×10^{-5}, only slightly higher than the glucose-only model. This shows that adding BMI slightly increases bias and variance but keeps the glucose coefficient stable and reliable overall.

Jackknife Glucose w/ Insulin Statistic:

Jackknife Mean Coef (Glucose w Insulin): 0.03888548

Jackknife Bias Estimate (Glucose w Insulin): 0.0003151924

Jackknife Variance Estimate (Glucose w Insulin): 1.553018e-05

Similarly, the glucose with insulin shows a very identical set of values, which confirms stability. With low bias and a slightly increased variance, it indicates that while the glucose coefficient stays stable, adding Insulin introduces a bit more variability.

Jackknife Visualization

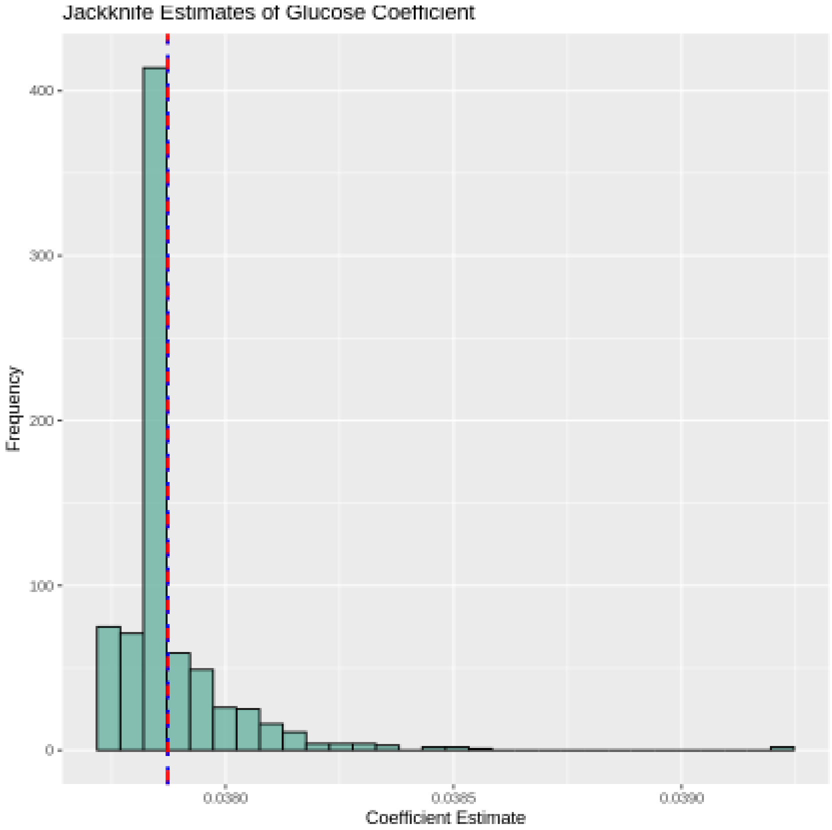


Figure 68. Jackknife Estimates of Glucose Coefficient

Looking at the figure above, it shows that there is an abundance of frequency at the 0.0375 which is clearly close with the original coefficient. Suggesting a proper variance along the line.

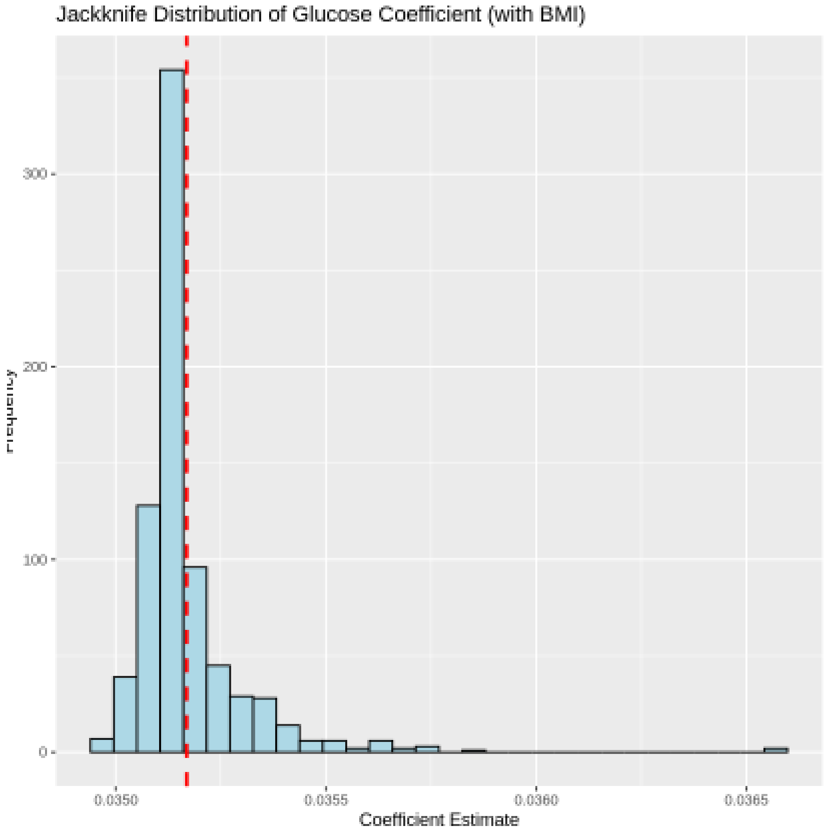


Figure 69. Jackknife Estimates of Glucose Coefficient with BMI

Similarly, the figure 69 shows a similar idea to the figure before, further explaining the numerical conclusion from the previous parts, that most data plays around their respective coefficients.

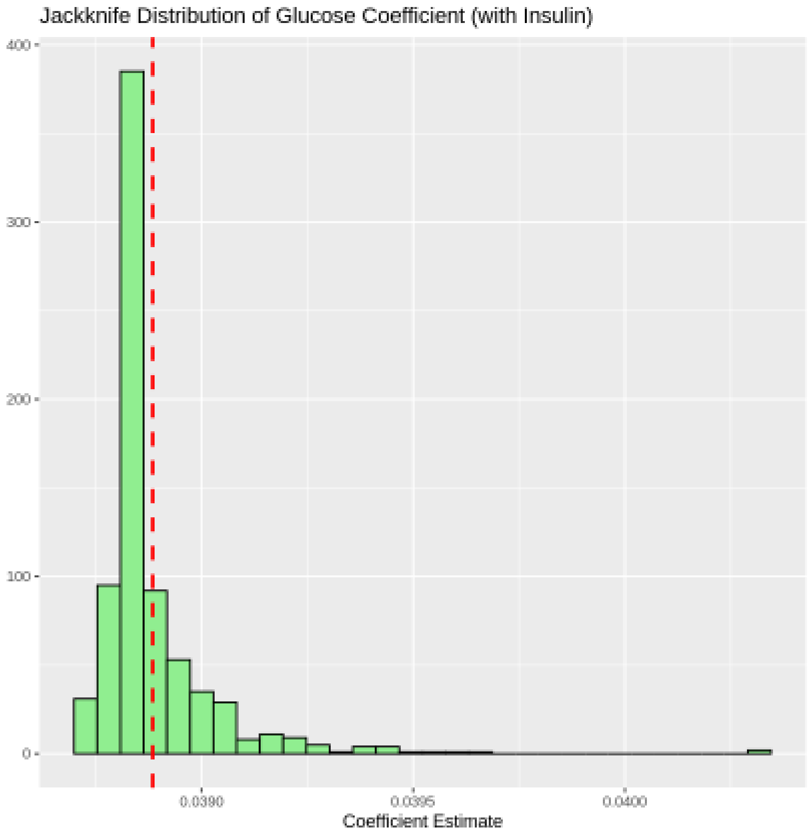


Figure 70. Jackknife Estimates of Glucose Coefficient with Insulin

Lastly, the figure 70 shows the jackknife distribution. Based on the visualization, among the three, this model shows the highest distance from the peak of the model to the coefficient, which could be a limitation for the number of bins present.

Theoretical Estimates

Original (theoretical) Coefficient Estimates

Glucose only: 0.03787304

Glucose + BMI: 0.03516896

Glucose + Insulin: 0.03888507

Looking at the theoretical coefficient estimates, the three almost share a similar estimate with differences only at the thousandths. The glucose + bmi produces the least value of the original coefficient while with insulin, it procures the highest, changing variability based on what the model is included.

Jackknife Standard Errors

Glucose only: 0.0001323647

Glucose + BMI: 0.0001336342

Glucose + Insulin: 0.000142388

Similarly, the standard error for all are on an acceptable margin with relatively low values, meaning that the variance of each would not stray away from the theoretical data.

Jackknife 95% Confidence Intervals (Percentile Method)

Glucose only: [0.03775619 - 0.03821105]

Glucose + BMI: [0.03502671 - 0.03550107]

Glucose + Insulin: [0.0387456 - 0.03925981]

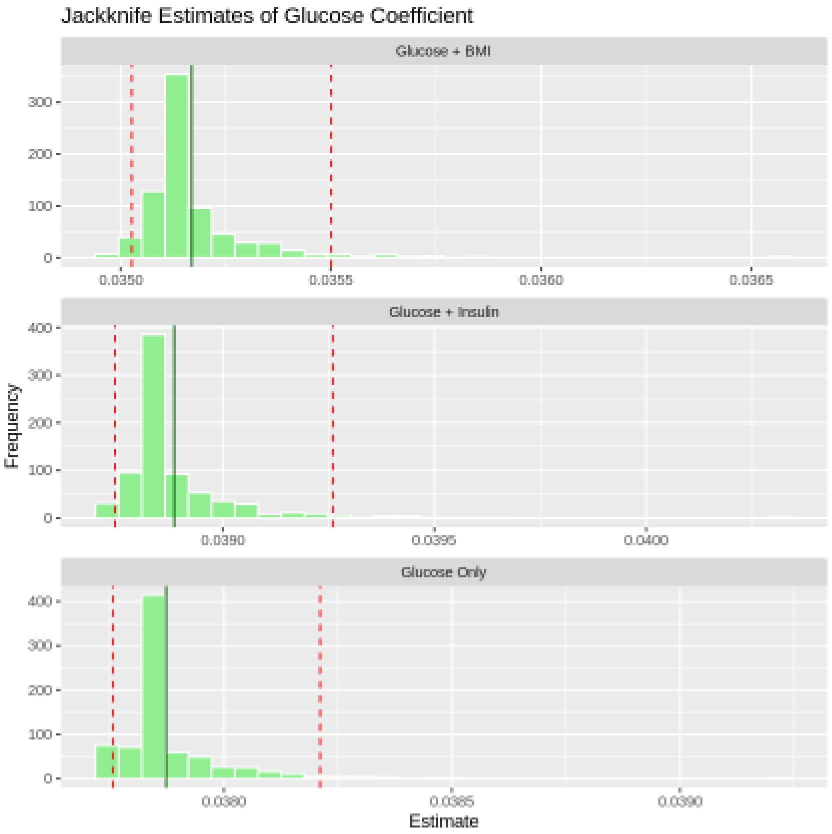


Figure 71-73 Jackknife Estimates of Glucose Coefficient

To better compare the values of the jackknife estimates between the three, this shows the similar plots but compared to each other side to side.

Resampling for Model Validation using Bootstrap

To evaluate the predictive capability of the different logistic regression models, performance metrics were assessed using bootstrap resampling. In this part, it includes the glucose only, and those with the iterations of with bmi, insulin, and with both.

Bootstrap Model Performance (Glucose only)

Average Accuracy: 0.7432706

Average Precision: 0.6905673

Average Recall: 0.4862307

Average F1 Score: 0.567582

Average RMSE: 0.4175247

Across 10,000 bootstrap iterations, the model showed an average accuracy of 74.33%, indicating that it correctly classified diabetes status in approximately three out of four cases.

The precision averaged 69.06%, meaning when the model predicted a positive diabetes outcome, it was correct most of the time. However, the recall was lower at 48.62%, indicating the model missed a considerable portion of actual diabetes cases. The F1 score—which balances precision and recall—was 56.76%, reflecting moderate predictive balance.

Finally, the Root Mean Squared Error (RMSE) averaged 0.418, suggesting moderate prediction error. Overall, while the model performs fairly well in identifying non-diabetic cases, its ability to correctly identify diabetic cases is limited when using glucose alone as the predictor.

For better utilization for future works, a function is introduced that allows the techniques to undergo a loop with 10,000 resamplings. To further assess and compare each of the models.

Model Performance Comparison (Bootstrap Resampling)

Model	Accuracy	Precision	Recall	F1 Score	RMSE
Glucose	0.7433	0.6906	0.4862	0.5676	0.4175
Glucose + BMI	0.7632	0.7198	0.5313	0.6088	0.4065
Glucose + Insulin	0.7417	0.6889	0.4816	0.5642	0.4175
Glucose + BMI + Insulin	0.7585	0.7105	0.5276	0.6027	0.4076

Table 13. Performance Comparison using Bootstrap Resampling on different Models

With the function given earlier, the similar bootstrap resampling for model validation can be introduced as well. Looking at the comparison of value, the Glucose-only model achieved an average accuracy of 74.33% but had relatively lower recall (48.62%) and F1 score (56.76%), indicating limited sensitivity in detecting diabetic cases.

Adding BMI improved overall performance, with the Glucose + BMI model reaching the highest accuracy (76.32%), precision (71.98%), recall (53.13%), and F1 score (60.88%), alongside a lower RMSE (0.4065), suggesting more balanced and accurate predictions.

The Glucose + Insulin model showed similar performance to the Glucose-only model, with no substantial gains in accuracy (74.17%) or recall (48.16%), indicating that Insulin alone did not notably enhance predictive ability.

The Glucose + BMI + Insulin model provided a slight improvement over the Glucose-only and Glucose + Insulin models, with accuracy (75.85%), recall (52.76%), and F1 score (60.27%) improving modestly, though it performed slightly below the Glucose + BMI model.

Overall, after assessment of different models in addition to the glucose, it shows that incorporating BMI alongside glucose performs the best in terms of model metrics. Because of the integration of BMI, it improves both the accuracy and balance within the precision and recall. Although all are similar in fashion and are near each other, it is important to note that at 76.32% accuracy means that this model provides more than 3/4 correct in diabetes predicting, which although is a mere upgrade to the glucose only, still is a better predictor when tackling issues such as health.

Conclusion

Based on the overall comprehensive statistical analysis, several conclusions can be drawn regarding the primary predictors of diabetes risk among women in pregnancy. In both the permutation testing and Bayesian inference, it was consistently identified as the main driving factor in the probability of diabetes, with BMI serving as an important secondary contributor. This conclusion is further supported by the bootstrap and jackknife resampling analyses, which confirmed the stability and low bias of the glucose coefficient, especially when combined with BMI, indicating a strong and consistent relationship.

For the distribution of key variables, exploratory data analysis revealed several predictors, e.g. Glucose, BMI, and Insulin, showed skewed distribution. Further backed by the KS tests. By this, it thus shows that the binary outcome variable aligned with a binomial distribution, validating the application of logistic regression for classification modeling.

In finalizing the report, the analyses established that through effective resampling techniques, it shows that the Glucose + BMI model consistently demonstrated as the best predictive model, reaching up to 76.32% accuracy rating. And with better evaluation to accuracy, recall, and F1-scores, it offers the most balanced prediction among all the models. Based on the logistic regression models, each one-unit increase in Glucose is associated with approximately a 3.7% increase in the odds of developing diabetes, as indicated by the coefficient (~0.038). This effect was consistent across resampling methods and remained the strongest predictor of diabetes risk.

In summary, the analysis establishes that Glucose is the strongest and most consistent predictor of diabetes risk, with BMI improving model performance without adding significant complexity. Through effective resampling techniques, it helped reduce the overfitting and addressed a concern of having a smaller dataset. These conclusions directly address the research questions by identifying key predictors, confirming variable distributions, and validating model performance through resampling strategies.

Future Exploration Ideas

To address the limitations of the study, future research could consider the following:

- **Data Imputation Techniques.** To replace the missing values from the dataset into values that represent the entire dataset for better representation and analysis

- **Broader dataset.** Since the current study only focused on female & pregnant patients on Pima Indians, future research can broaden the dataset that include other participants such as other genders, ethnicities and groups.

Model Validation. This is to cross-validate the preliminary model used to create a robust and accurate representation of the dataset.