

University of San Carlos
DCISM

Final Culminating Activity

MATH 3109

PREPARED BY:
Woodrow A. Fajardo

SUBMITTED TO:
Prof. Elmer S. Poliquit

MATH-3109
NOVEMBER 2025

Contents

1	Introduction	1
1.1	Project Objective	1
1.2	Literature Review	1
1.3	Raw Data	1
2	Data Cleaning	2
2.1	Converting Zeroes Values to NA	2
2.2	MICE Imputation	3
2.3	Imputation Validation	3
3	Exploratory Analysis	7
3.1	Correlation Matrix	7
3.2	Analyzing Linearity	7
4	Modelling Logistic Regression	8
4.1	Model Specification	8
4.2	Odds Ratio Interpretation	8
4.3	Model Performance	9
5	Linear Regression (Predict Insulin)	9
5.1	B-Spline Justification	10
5.2	Comparison Linear vs. B-Spline	11
5.3	Residual Diagnostics	12
6	Discussions	12
6.1	Validation of Literature Claims	12
6.2	Key Findings	13
6.3	Limitations (if any)	13
7	Appendix	13

Raw Data based from Kaggle:

1 Introduction

Diabetes is a significant public health concern that demands accurate methods for early detection. Predictive modeling allows us to identify individuals at high risk before severe complications arise. This study utilizes the Pima Indians Diabetes Database to construct robust statistical models. We aim to determine which diagnostic measurements most accurately predict the onset of diabetes and to model the non-linear fluctuations of insulin levels using applied regression techniques.

1.1 Project Objective

This project aims to predict the onset of diabetes based on diagnostic measures. We will also model insulin levels using regression techniques. The goal is to identify key risk factors and understand non-linear relationships in the medical data.

1.2 Literature Review

Nassiwa and Zeng identify Glucose and BMI as the primary predictors of diabetes. This study seeks to replicate those findings using the Pima Indians dataset. We hypothesize that metabolic markers will outrank demographic factors like age in predictive power.

1.3 Raw Data

The dataset consists of 768 observations from the Pima Indians Diabetes Database. It includes medical predictor variables and one target variable, Outcome. Initial inspection reveals missing values encoded as zeros, which requires correction before analysis.

R | SETTING UP ENVIRONMENT

```
1 library(mice)
2 library(corrplot)
3 library(ggplot2)
4 library(splines)
5 library(performance)
6 library(see)
7 library(ggpubr)
8
9 setwd("~/School/MATH-3109/FINAL-1")
10 raw_data <- read.csv("raw_dataset.csv")
11 View(raw_data)
12 head(raw_data)
```

RAW DATA HEAD

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
6	148	72	35	0
1	85	66	29	0
8	183	64	0	0
1	89	66	23	94
0	137	40	35	168
5	116	74	0	0

BMI	DiabetesPedigreeFunction	Age	Outcome
33.6	0.627	50	1
26.6	0.351	31	0
23.3	0.672	32	1
28.1	0.167	21	0
43.1	2.288	33	1
25.6	0.201	30	0

2 Data Cleaning

Real-world medical data is rarely perfect. We must first address biologically impossible zero values that represent missing data to avoid biased results.

2.1 Converting Zeroes Values to NA

The summary confirms substantial missingness disguised as zeros.

R | GETTING COLUMNS WITH ZERO VALUES

```
1 as.matrix(colSums(raw_data == 0))
```

COLUMNS AND NUMBER OF ZERO VALES

Pregnancies	111
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	500

R | CONVERTING ZERO VALUES TO NA

```
1 col_w_zeroes <- c("Pregnancies", "Glucose", "BloodPressure", "SkinThickness",
2                   "Insulin", "BMI")
3 raw_data[col_w_zeroes] <- lapply(raw_data[col_w_zeroes], function(x)
4                                ifelse(x == 0, NA, x))
5
6 summary(raw_data)
```

VISUALIZING SUMMARY ACCOUNTING NA's

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
Min. : 1.000	Min. : 44.0	Min. : 24.00	Min. : 7.00	Min. : 14.00
1st Qu.: 2.000	1st Qu.: 99.0	1st Qu.: 64.00	1st Qu.:22.00	1st Qu.: 76.25
Median : 4.000	Median :117.0	Median : 72.00	Median :29.00	Median :125.00
Mean : 4.495	Mean :121.7	Mean : 72.41	Mean :29.15	Mean :155.55
3rd Qu.: 7.000	3rd Qu.:141.0	3rd Qu.: 80.00	3rd Qu.:36.00	3rd Qu.:190.00
Max. :17.000	Max. :199.0	Max. :122.00	Max. :99.00	Max. :846.00
NA's :111	NA's :5	NA's :35	NA's :227	NA's :374

BMI	DiabetesPedigreeFunction	Age	Outcome
Min. :18.20	Min. :0.0780	Min. :21.00	Min. :0.000
1st Qu.:27.50	1st Qu.:0.2437	1st Qu.:24.00	1st Qu.:0.000
Median :32.30	Median :0.3725	Median :29.00	Median :0.000
Mean :32.46	Mean :0.4719	Mean :33.24	Mean :0.349
3rd Qu.:36.60	3rd Qu.:0.6262	3rd Qu.:41.00	3rd Qu.:1.000
Max. :67.10	Max. :2.4200	Max. :81.00	Max. :1.000
NA's :11			

Insulin has the highest missing rate, with 374 zero entries. Variables like Glucose, Blood Pressure, and BMI also contain invalid zeros that require correction.

2.2 MICE Imputation

Dropping rows with missing data would lose too much information. We will use Multivariate Imputation by Chained Equations (MICE) with predictive mean matching to estimate and fill these missing values.

R | MICE IMPUTATION PMM as METHOD

```
1 init <- mice(raw_data, maxit = 0)
2 methods <- init$method #$$
3 methods[col_w zeroes] <- "pmm"
4 imputed_data <- mice(raw_data, method = methods, m = 5, maxit = 5, seed = 123)
5 final_data <- complete(imputed_data, 1)
6 head(final_data)
```

View full R script output here.

R | CHECK FINAL DATA CONTAINS NA's

```
1 anyNA(final_data)
```

RETURN TRUE CONTAINS NA's OTHERWISE NONE

```
[1] FALSE
```

2.3 Imputation Validation

We must verify that our statistical imputation did not distort the natural distribution of the data. Visual inspection confirms the integrity of the data. The imputed distributions (blue) closely mirror the original distributions (red). This suggests the MICE algorithm preserved the underlying data structure without introducing bias.

Refer to Appendix Section for the full R script.

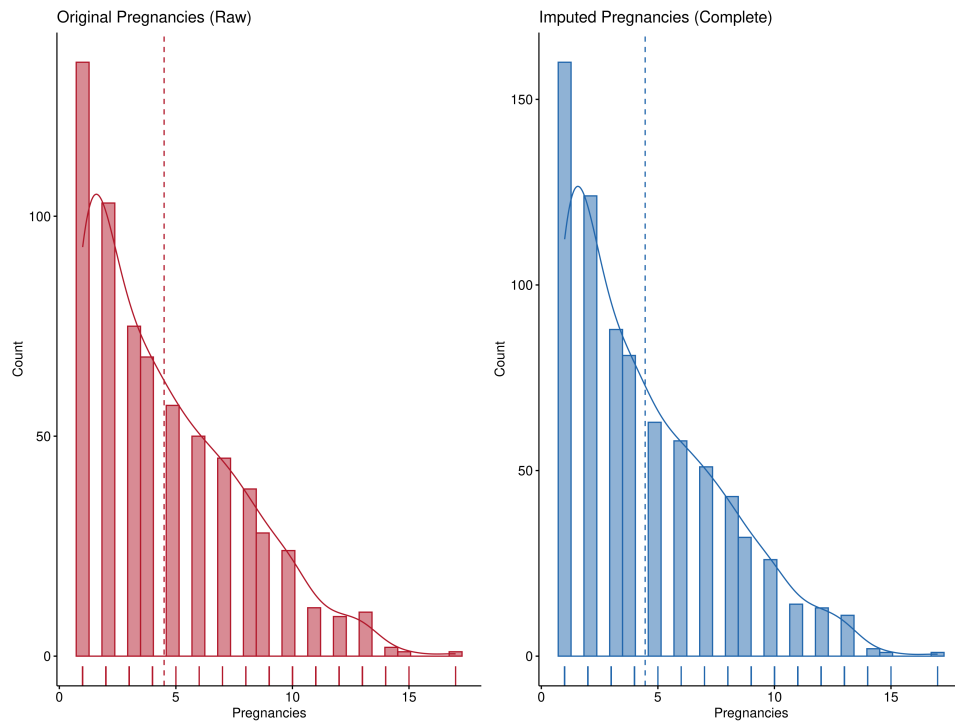


FIGURE 1
Before and After Imputation (Pregnancies).

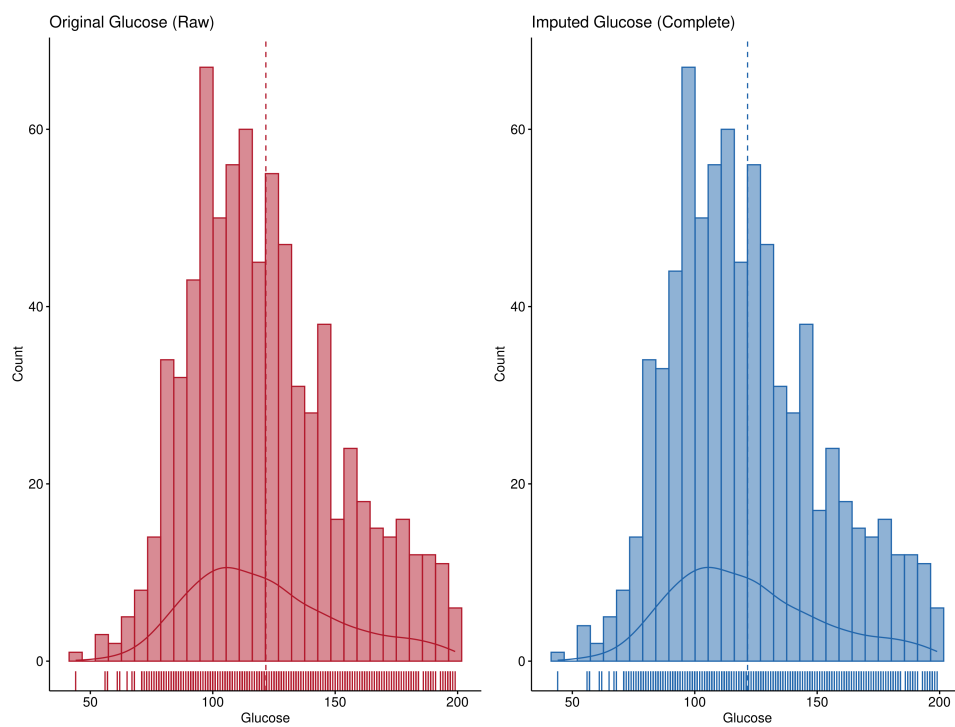


FIGURE 2
Before and After Imputation (Glucose).

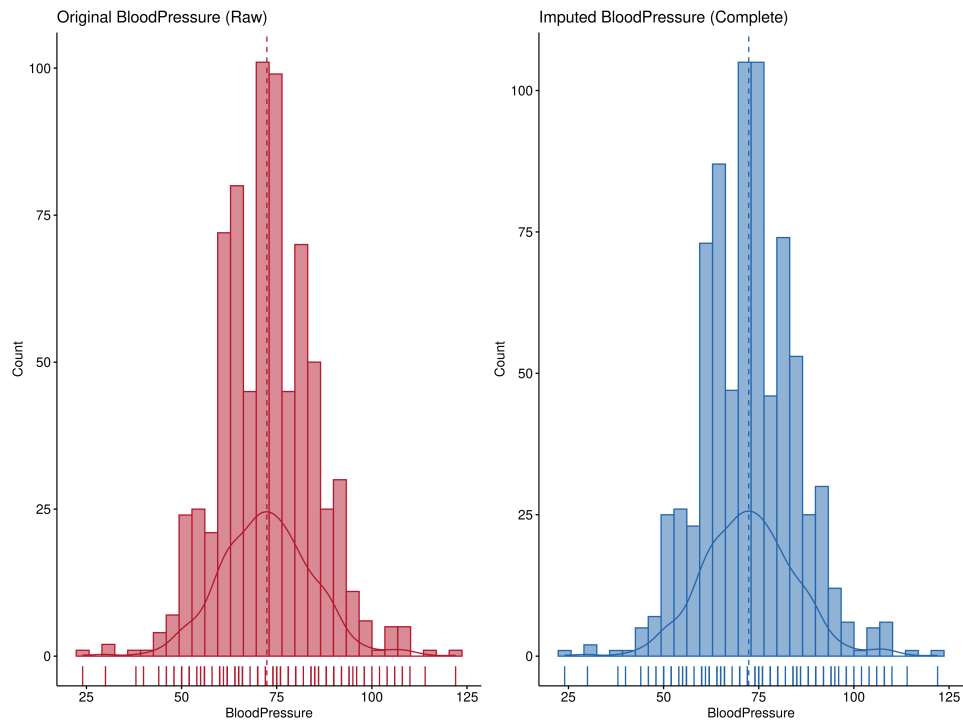


FIGURE 3
Before and After Imputation (BloodPressure).

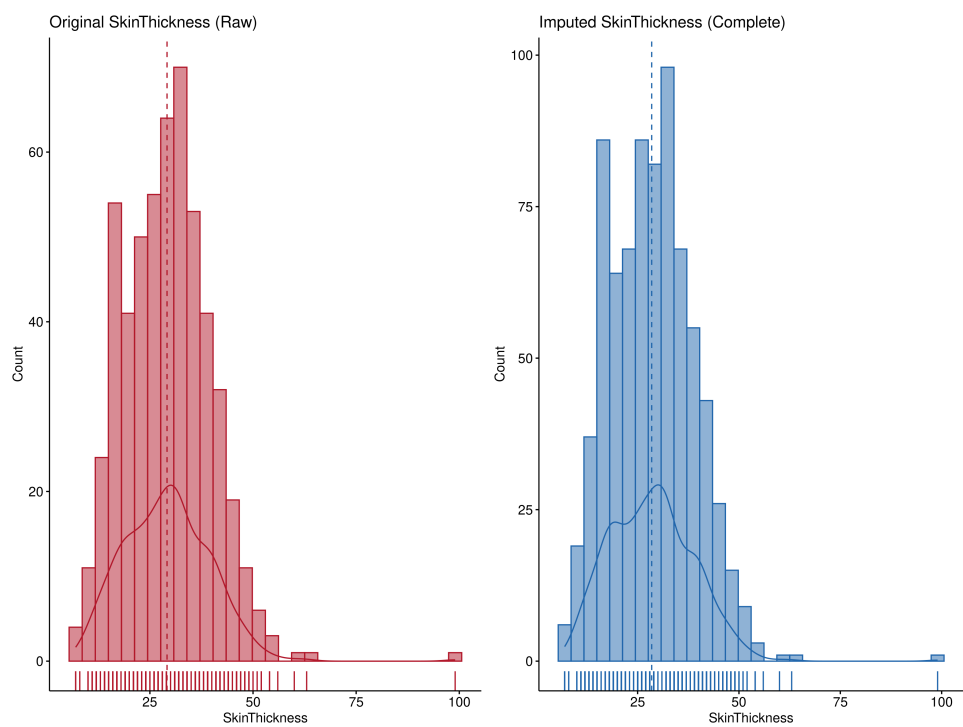


FIGURE 4
Before and After Imputation (SkinThickness).

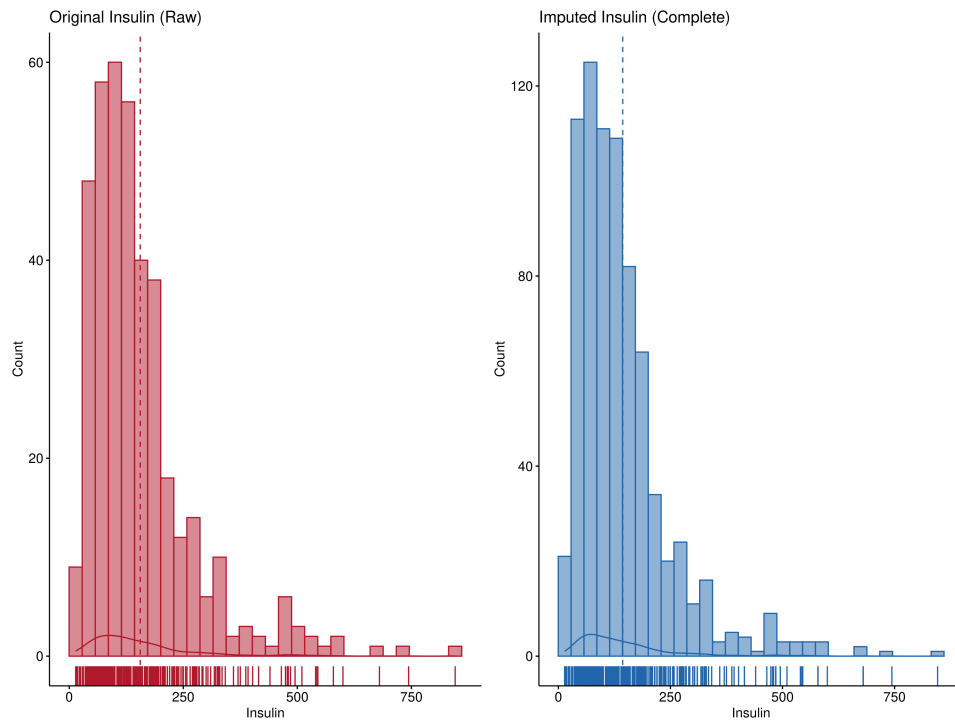


FIGURE 5
Before and After Imputation (Insulin).

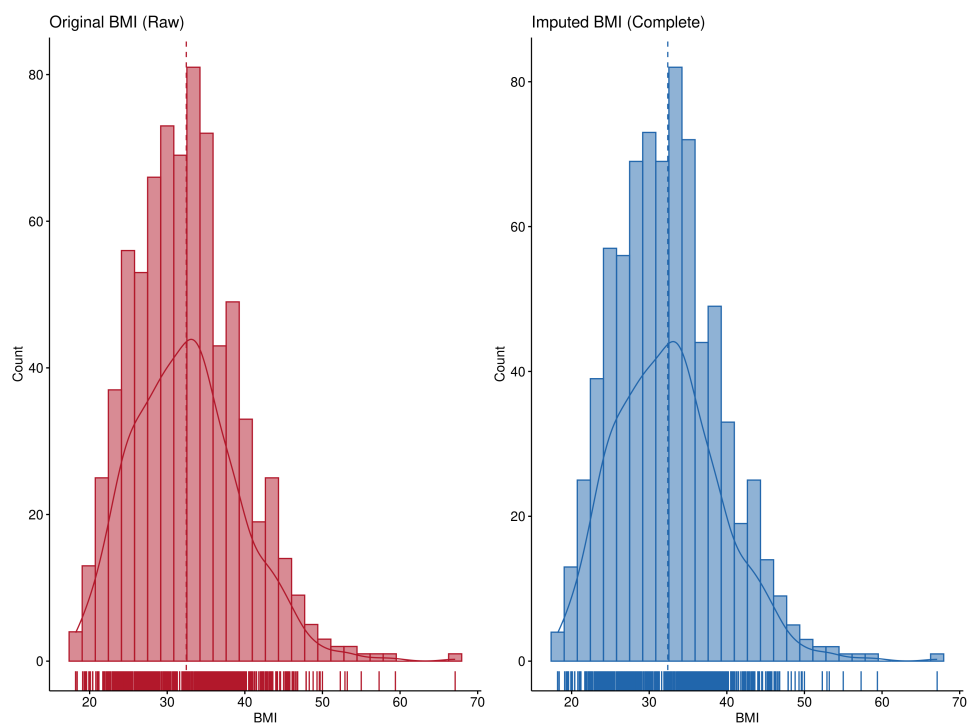


FIGURE 6
Before and After Imputation (BMI).

3 Exploratory Analysis

Before modeling, we examine the relationships between variables to identify potential predictors. We specifically check for multicollinearity to ensure our regression models remain stable and interpretable.

3.1 Correlation Matrix

The matrix reveals strong correlations between **Glucose** and the diabetes **Outcome**. We also observe multicollinearity between **Age** and **Pregnancies**. This validates our selection of predictors but suggests caution regarding variance inflation in the models.

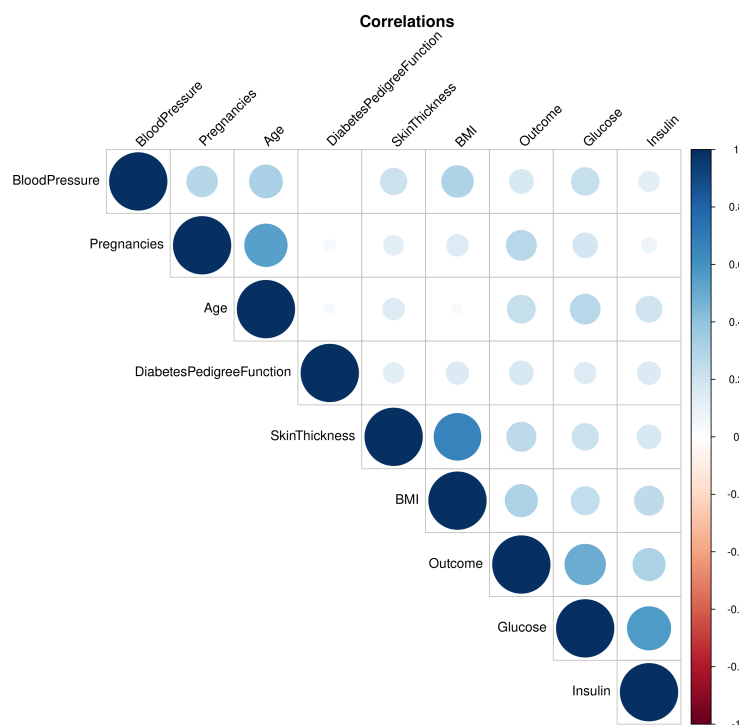


FIGURE 7
Correlation Matrix Plot

STATISTICAL MODELING

```
1 cor_matrix <- cor(final_data, use = "complete.obs")
2 corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust",
3          tl.col = "black", tl.srt = 45, title = "Correlations", mar=c(0,0,1,0))
```

3.2 Analyzing Linearity

We test the linearity assumption for our continuous regression model by examining the relationship between Age and Insulin. The local regression plot shows a distinct curve. Insulin levels rise in early adulthood but plateau after age 50. A simple linear line cannot capture this, justifying the need for non-linear modeling techniques.

4 Modelling Logistic Regression

We fit a logistic regression model to predict the binary diabetes outcome using our cleaned dataset. The model separates diabetic and non-diabetic patients based on the selected risk factors.

4.1 Model Specification

Glucose, BMI, and Pregnancies are highly significant predictors ($p < 0.001$). However, Age is not statistically significant ($p = 0.337$). This is likely due to its strong correlation with Pregnancies, which absorbed the predictive power in this model.

STATISTICAL MODELING

```
1 set.seed(123)
2 sample_index <- sample(1:nrow(final_data), 0.8 * nrow(final_data))
3 train_data <- final_data[sample_index, ]
4 test_data <- final_data[-sample_index, ]
5
6 log_model <- glm(Outcome ~ Glucose + BMI + Age + Pregnancies +
  ↪ DiabetesPedigreeFunction,
7               data = train_data, family= "binomial")
8
9 summary(log_model)
```

TERMINAL OUTPUT

```
glm(formula = Outcome ~ Glucose + BMI + Age + Pregnancies +
  ↪ DiabetesPedigreeFunction,
    family = "binomial", data = train_data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -9.200623   0.809675 -11.363  < 2e-16 ***
Glucose         0.037413   0.004099   9.128  < 2e-16 ***
BMI             0.079814   0.016719   4.774 1.81e-06 ***
Age             0.009594   0.010003   0.959 0.337506
Pregnancies     0.130485   0.037653   3.465 0.000529 ***
DiabetesPedigreeFunction 0.659625   0.321580   2.051 0.040247 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 796.42  on 613  degrees of freedom
Residual deviance: 568.82  on 608  degrees of freedom
AIC: 580.82

Number of Fisher Scoring iterations: 5
```

4.2 Odds Ratio Interpretation

We convert the model coefficients into odds ratios to quantify the biological risk in human-readable terms. For every one-unit increase in BMI, the odds of diabetes rise by roughly 8.3%. Glucose shows a similar positive risk. The Diabetes Pedigree Function has the highest ratio, nearly doubling the odds for each unit increase.

STATISTICAL MODELING

```
1 exp(coef(log_model))
```

TERMINAL OUTPUT

(Intercept)	Glucose	BMI
↪ Age	Pregnancies DiabetesPedigreeFunction	
0.0001009765	1.0381215630	1.0830855110
↪ 1.0096399484	1.1393803946	1.9340669910

4.3 Model Performance

We assess the model's goodness-of-fit by examining the deviance and AIC values. The model significantly reduces deviance from 796 to 568. This drop indicates that our selected predictors explain a substantial amount of the variation in diabetes onset.

5 Linear Regression (Predict Insulin)

We shift to a continuous model to predict Insulin levels based on Glucose, BMI, and Age. This analysis aims to capture the factors driving insulin fluctuation.

STATISTICAL MODELING

```
1 linear_simple <- lm(Insulin ~ Glucose + BMI + Age, data = final_data)
2 summary(linear_simple)
```

TERMINAL OUTPUT

```
Call:
lm(formula = Insulin ~ Glucose + BMI + Age, data = final_data)

Residuals:
    Min       1Q   Median       3Q      Max
-289.02  -45.91  -14.70   25.01  567.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -167.6456    19.1120  -8.772  < 2e-16 ***
Glucose       1.8543     0.1125  16.481  < 2e-16 ***
BMI           2.0651     0.4791   4.310 1.84e-05 ***
Age           0.5693     0.2841   2.004  0.0455 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.01 on 764 degrees of freedom
Multiple R-squared:  0.3401,    Adjusted R-squared:  0.3375
F-statistic: 131.2 on 3 and 764 DF,  p-value: < 2.2e-16
```

STATISTICAL MODELING

```
1 linear_spline <- lm(Insulin ~ Glucose + BMI + bs(Age, degree = 3), data =
  ↪ final_data)
2 summary(linear_spline)
```

Statement

TERMINAL OUTPUT

```

Call:
lm(formula = Insulin ~ Glucose + BMI + bs(Age, degree = 3), data = final_data)

Residuals:
    Min       1Q   Median       3Q      Max
-299.47  -47.78  -14.42   26.00  554.90

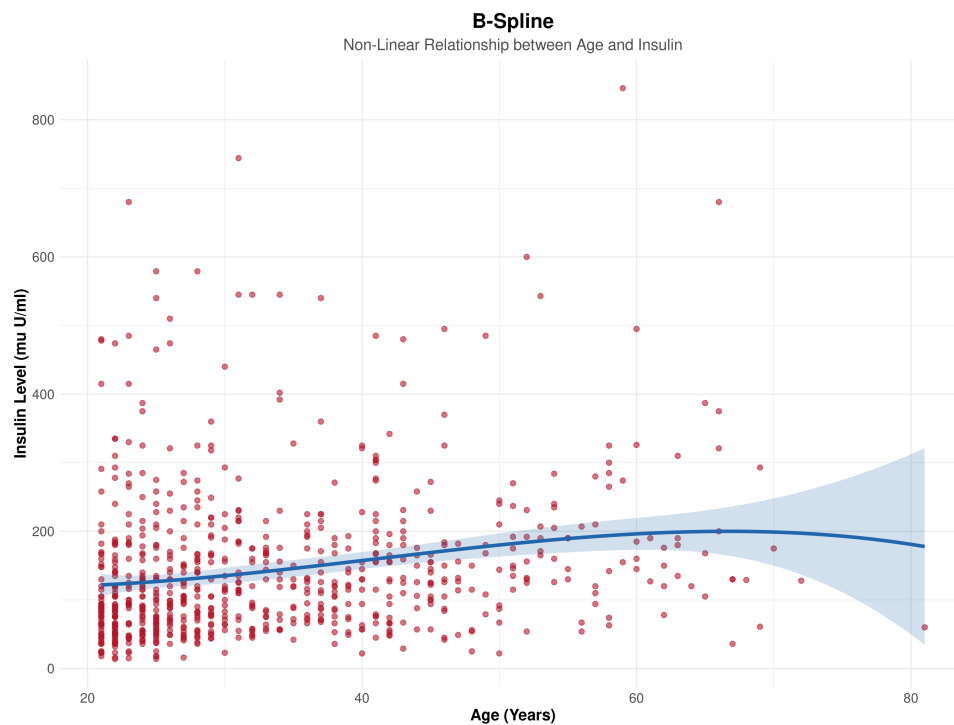
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -151.3071    18.2230  -8.303 4.61e-16 ***
Glucose         1.8612     0.1123  16.571 < 2e-16 ***
BMI             2.2633     0.4871   4.646 3.98e-06 ***
bs(Age, degree = 3)1 -57.2504    32.8240  -1.744  0.0815 .
bs(Age, degree = 3)2  80.9994    50.5390   1.603  0.1094
bs(Age, degree = 3)3   5.4944    62.7432   0.088  0.9302
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 88.82 on 762 degrees of freedom
Multiple R-squared:  0.3447,    Adjusted R-squared:  0.3404
F-statistic: 80.16 on 5 and 762 DF,  p-value: < 2.2e-16

```

Statement**5.1 B-Spline Justification**

Standard linear regression assumes a constant rate of change. Our EDA showed that aging effects saturate over time. Therefore, we apply a cubic B-spline to Age to model this biological non-linearity. Refer to Appendix Section for the full R script.

**FIGURE 8**

A standard centered figure.

5.2 Comparison Linear vs. B-Spline

We statistically compare the simple linear model against the B-spline model using an ANOVA test. The ANOVA results show a p-value of 0.07. While strictly above the 0.05 threshold, it suggests marginal significance. Given the visual evidence of curvature, we retain the spline model for its biological realism.

STATISTICAL MODELING

```
1 linear_simple <- lm(Insulin ~ Glucose + BMI + Age, data = final_data)
2 linear_spline <- lm (Insulin ~ Glucose + BMI + bs(Age, degree = 3), data =
  ↪ final_data)
3 anova(linear_simple, linear_spline)
```

TERMINAL OUTPUT

Analysis of Variance Table

Model 1: Insulin ~ Glucose + BMI + Age

Model 2: Insulin ~ Glucose + BMI + bs(Age, degree = 3)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	764	6052997				
2	762	6010880	2	42116	2.6696	0.06993

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Statement

5.3 Residual Diagnostics

We examine the residuals to ensure the model assumptions of homoscedasticity and normality are met. Diagnostic plots show that the residuals are randomly scattered, satisfying the linearity assumption. Point 460 is flagged as influential but remains within acceptable Cook's distance limits. The model fits the data well.

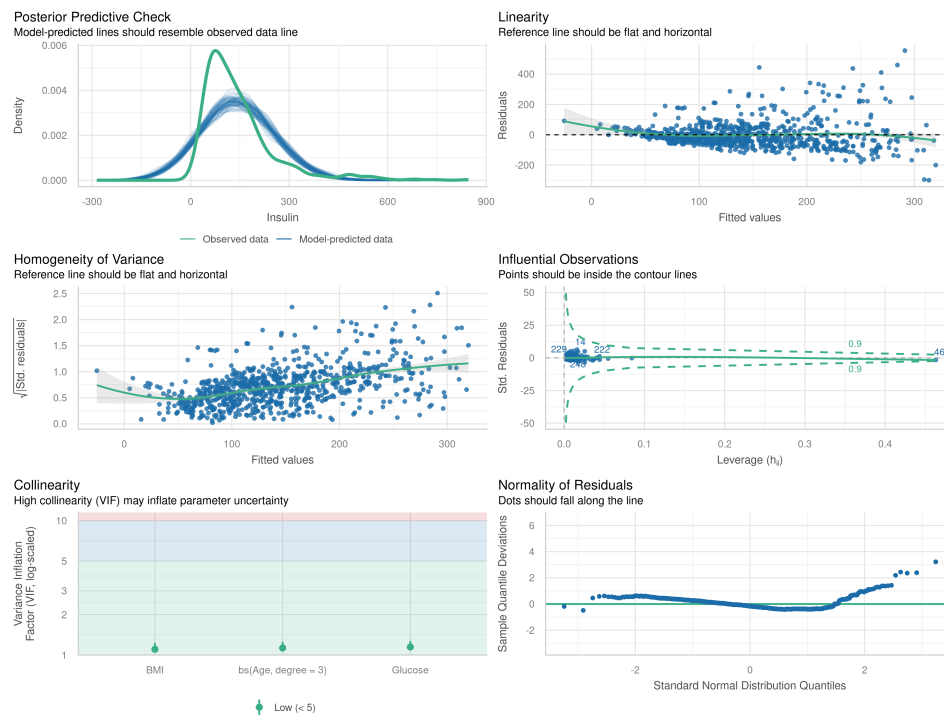


FIGURE 9

A standard centered figure.

STATISTICAL MODELING

```
1 image_rend <- check_model(linear_spline)
2 png("splien_model.png", width = 12, height = 9, units = "in", res = 300)
3 print(image_rend)
4 dev.off()
```

Statement

6 Discussions

We synthesize our findings to validate our initial hypotheses. This discussion connects our statistical outputs back to the medical literature and the project objectives.

6.1 Validation of Literature Claims

Our results confirm the claims of Nassiwa and Zeng. Glucose and BMI were the strongest predictors in our logistic model ($p < 2e - 16$). This proves that immediate metabolic markers are more predictive than general demographics like age.

6.2 Key Findings

This study produced three major statistical insights. We successfully imputed missing data without bias. Logistic regression identified Glucose, BMI, and Pregnancies as primary risk factors. Finally, we demonstrated that the relationship between Age and Insulin is non-linear, requiring spline modeling.

6.3 Limitations (if any)

The dataset has inherent constraints despite our cleaning efforts. The dataset is relatively small, with only 768 observations. Despite robust imputation, the original missingness in Insulin was high. Future studies would benefit from a larger, more complete dataset to improve precision.

7 Appendix

HISTOGRAM COMPARISON CODE EXAMPLE

```

1 p1 <- gg_histogram(raw_data, x = "Pregnancies",
2   title = "Original Pregnancies (Raw)",
3   xlab = "Pregnancies", ylab = "Count",
4   fill = "#b2182b",
5   color = "#b2182b",
6   add = "mean", rug = TRUE, add_density = TRUE,
7   ggtheme = theme_pubr()
8 )
9
10 p2 <- gg_histogram(final_data, x = "Pregnancies",
11   title = "Imputed Pregnancies (Complete)",
12   xlab = "Pregnancies", ylab = "Count",
13   fill = "#2166ac",
14   color = "#2166ac",
15   add = "mean", rug = TRUE, add_density = TRUE,
16   ggtheme = theme_pubr()
17 )
18
19 ggarrange(p1, p2, ncol = 2, nrow = 1)

```

STATISTICAL MODELING

```

1 p1 <- gghistogram(raw_data, x = "Glucose",
2   title = "Original Glucose (Raw)",
3   xlab = "Glucose", ylab = "Count",
4   fill = "#b2182b",
5   color = "#b2182b",
6   add = "mean", rug = TRUE, add_density = TRUE,
7   ggtheme = theme_pubr()
8 )
9
10 p2 <- gghistogram(final_data, x = "Glucose",
11   title = "Imputed Glucose (Complete)",
12   xlab = "Glucose", ylab = "Count",
13   fill = "#2166ac",
14   color = "#2166ac",
15   add = "mean", rug = TRUE, add_density = TRUE,
16   ggtheme = theme_pubr()
17 )
18
19 ggarrange(p1, p2, ncol = 2, nrow = 1)

```

STATISTICAL MODELING

```

1 p1 <- gghistogram(raw_data, x = "BloodPressure",
2   title = "Original BloodPressure (Raw)",
3   xlab = "BloodPressure", ylab = "Count",
4   fill = "#b2182b",
5   color = "#b2182b",
6   add = "mean", rug = TRUE, add_density = TRUE,
7   ggtheme = theme_pubr()
8 )
9
10 p2 <- gghistogram(final_data, x = "BloodPressure",
11   title = "Imputed BloodPressure (Complete)",
12   xlab = "BloodPressure", ylab = "Count",
13   fill = "#2166ac",
14   color = "#2166ac",
15   add = "mean", rug = TRUE, add_density = TRUE,
16   ggtheme = theme_pubr()
17 )
18
19 ggarrange(p1, p2, ncol = 2, nrow = 1)

```


STATISTICAL MODELING

```

1 p1 <- gg_histogram(raw_data, x = "SkinThickness",
2   title = "Original SkinThickness (Raw)",
3   xlab = "SkinThickness", ylab = "Count",
4   fill = "#b2182b",
5   color = "#b2182b",
6   add = "mean", rug = TRUE, add_density = TRUE,
7   ggtheme = theme_pubr()
8 )
9
10 p2 <- gg_histogram(final_data, x = "SkinThickness",
11   title = "Imputed SkinThickness (Complete)",
12   xlab = "SkinThickness", ylab = "Count",
13   fill = "#2166ac",
14   color = "#2166ac",
15   add = "mean", rug = TRUE, add_density = TRUE,
16   ggtheme = theme_pubr()
17 )
18
19 ggarrange(p1, p2, ncol = 2, nrow = 1)

```

STATISTICAL MODELING

```

1 p1 <- gg_histogram(raw_data, x = "Insulin",
2   title = "Original Insulin (Raw)",
3   xlab = "Insulin", ylab = "Count",
4   fill = "#b2182b",
5   color = "#b2182b",
6   add = "mean", rug = TRUE, add_density = TRUE,
7   ggtheme = theme_pubr()
8 )
9
10 p2 <- gg_histogram(final_data, x = "Insulin",
11   title = "Imputed Insulin (Complete)",
12   xlab = "Insulin", ylab = "Count",
13   fill = "#2166ac",
14   color = "#2166ac",
15   add = "mean", rug = TRUE, add_density = TRUE,
16   ggtheme = theme_pubr()
17 )
18
19 ggarrange(p1, p2, ncol = 2, nrow = 1)

```

STATISTICAL MODELING

```

1 p1 <- gghistogram(raw_data, x = "BMI",
2   title = "Original BMI (Raw)",
3   xlab = "BMI", ylab = "Count",
4   fill = "#b2182b",
5   color = "#b2182b",
6   add = "mean", rug = TRUE, add_density = TRUE,
7   ggtheme = theme_pubr()
8 )
9
10 p2 <- gghistogram(final_data, x = "BMI",
11   title = "Imputed BMI (Complete)",
12   xlab = "BMI", ylab = "Count",
13   fill = "#2166ac",
14   color = "#2166ac",
15   add = "mean", rug = TRUE, add_density = TRUE,
16   ggtheme = theme_pubr()
17 )
18
19 ggarrange(p1, p2, ncol = 2, nrow = 1)

```

STATISTICAL MODELING

```

1 ggplot(final_data, aes(x = Age, y = Insulin)) +
2   geom_point(color = "#b2182b", alpha = 0.6, size = 2) +
3   geom_smooth(method = "lm",
4     formula = y ~ bs(x, degree = 3),
5     color = "#2166ac",
6     fill = "#2166ac",
7     alpha = 0.2,
8     size = 1.5) +
9
10   labs(title = "B-Spline",
11     subtitle = "Non-Linear Relationship between Age and Insulin",
12     x = "Age (Years)",
13     y = "Insulin Level (mu U/ml)") +
14
15   theme_minimal() +
16   theme(
17     plot.title = element_text(face = "bold", hjust = 0.5, size = 18),
18     plot.subtitle = element_text(hjust = 0.5, color = "gray30", size = 14),
19     axis.title = element_text(face = "bold", size = 14),
20     axis.text = element_text(size = 12))

```