Part A

In this report I will discuss the development and evaluation of a logistic regression model to predict the presence of diabetes. Data analysis was completed using R (version 4.4.1). Table 1.1 provides a descriptive presentation of the dataset stratified by diabetes status.

Table 1.1. Descriptive Statistics

Variable	Definition		No Diabetes N = 1816 (66.6%)	Diabetes N = 952 (34.4%)	*p*-value
Pregnancy	Number of pregnancies	Count (%)			< 0.001
0			274 (15.1%)	138 (14.5%)	
1			383 (21.2%)	108 (11.3%)	
2			554 (30.5%)	210 (21.1%)	
3			389 (21.4%)	363 (38.1%)	
4			200 (11.0%)	139 (14.6%)	
5			16 (0.9%)	3 (0.3%)	
Glucose	Concentration of blood glucose[1] (mg / dL)	Med (min-max)	107.00 (44.00-197.00)	141.00 (78.00-199.00)	< 0.001
		IQR	31.25	47.00	
Blood Pressure	Diastolic blood pressure (mmHg)	Med (min-max)	70.00 (24.00-122.00)	74.00 (30.00-114.00)	< 0.001
		IQR	16.00	14.00	
Skin thickness	Triceps skinfold thickness (mm)	Med (min-max)	27.00 (7.00-110.00)	32.00 (7.00-99.00)	< 0.001
		IQR	9.00	6.00	
BMI	Body mass index (kg / m^2)	Med (min-max)	30.10 (18.20-80.60)	34.30 (20.10-67.10)	< 0.001
		IQR	9.72	7.83	
Diabetes Genetic Score	Genetic score for diabetes	Med (min-max)	0.34 (0.08-2.33)	0.44 (0.09-2.42)	< 0.001
		IQR	0.34	0.47	
Age	Age (years)	Med (min-max)	26.00 (21.00-81.00)	36.00 (21.00-70.00)	< 0.001
		IQR	14.00	16.00	

Med = median
min = minimum
max = maximum
IQR = Inter quartile range

[1] Blood glucose concentration over two hours in oral glucose tolerance test (mg / dL)

Data Redundancy Implications

Only 803 of the 2768 rows in the dataset are unique. However, duplicated rows were not removed due to unique patient identifiers. High dataset redundancy may cause over-fitting for repetitive patterns. A 50:50 test:train split was used to minimise this.

Addressing Missing Data

48.05% of the observations have missing insulin values. An increase of goodness-of-fit was assessed as a significant reduction in AIC and increase in AUC. No missing data approaches showed significant improvement relative to the removal of insulin from the model. In the complete records analysis model, insulin was not a statistically significant predictor of diabetes (OR=1.00, 95%CI [1.00, 1.00]). Thus, insulin was not included as a model covariate and omitted from Table 1.1.

Logistic Regression Assumptions

Prior to investigating assumptions and model building, the data was randomly split into a 50:50 train-test split, to enable internal model validation.

The assumption of independence of errors could not be analysed without further information about the data collection methods. Moderate collinearity (0.54) between Body Mass Index (BMI) and triceps skinfold thickness was below the threshold of concern (0.7). Identification of high leverage points will be discussed relative to the final model.

To assess linearity in the logit, the predicted log-odds versus the individual continuous variables were plotted (Appendix Figure 2.1). Age, blood pressure, BMI and pregnancies did not display approximate linearity, indicating the need for variable transformations.

Variable Transformations

A basic model was created using the training data (AIC=1316.88, AUC=0.84). Number of pregnancies is not a continuous covariate, as indicated by the distinct group effect in the logit plot. Patients with 5 pregnancies (N=19) are under-represented in the dataset, informing the use of a "4+" category. The adjusted model displayed a moderate reduction in AIC (1308.28) and no change in AUC (0.84) with respect to the basic model.

Log-age was investigated to address the gradual decrease in influence evident in the logit plot. This reduced AIC (1304.52) and had no effect on AUC (0.84) relative to the basic model. A log-age logit plot showed that lower values were successfully spread, but the curve still plateaued at higher values (log-age > 4) (Appendix Figure 2.2). Age was then categorised ("18-29", "30-44", "45-59", "60+"),

resulting in a large reduction in AIC (1259.27) and an increase in AUC (0.86). Interpretability was also improved, making categorisation the preferred transformation.

Transformations of blood pressure explored included polynomials, splines and categorisation. None resulted in a significant improvement in AIC and/or AUC relative to the basic model.

BMI was categorised in line with clinically significant groups. The "Underweight" category (BMI<18.5) is under-represented in the dataset (N=14). To avoid introducing instability, the reference category of "Normal Weight" was set to BMI between 18 and 25. The resulting model showed significant reduction in AIC (1291.13) and a minor increase in AUC (0.85) relative to the basic model.

Interaction Terms

To correctly capture the relationship between variables, age, BMI and pregnancies were categorised in the reference model (AIC=1228.70, AUC=0.87) and interaction terms. Based on clinical rational the interaction terms explored were glucose with diabetes genetic score, BMI and age, and diabetes genetic score with BMI and age. Glucose x BMI resulted in the most significant reduction in AIC (1195.66) and increase in AUC (0.88) relative to the reference model. The likelihood ratio test showed that the inclusion of the interaction term significantly improved the model (p<0.05).

Model Design

The final model was designed iteratively via best subset selection as per Table 1.2. Pregnancies and blood pressure were removed from Model 3 as they were statistically insignificant at the p<0.05 level. The inclusion of insignificant variables introduces noise without contributing meaningfully to prediction. Model 4, which included the interaction term, displayed the lowest AIC and deviance, and the highest log-likelihood.

Table 1.2. Predictive models for diabetes risk factors

	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	-0.68 *** (0.06)	-9.03 *** (0.74)	-8.96 *** (0.58)	-13.38 *** (2.17)	-13.52 *** (2.17)
Pregnancies					
1		-0.14 (0.27)			
2		-0.00 (0.25)			
3		0.46 (0.26)			
4+		0.03 (0.28)			
Glucose		0.04 *** (0.00)	0.04 *** (0.00)	0.07 *** (0.01)	0.07 *** (0.01)
Blood Pressure		0.00 (0.01)			
Triceps Skinfold thickness		0.04 *** (0.01)	0.04 *** (0.01)	0.04 *** (0.01)	0.05 *** (0.01)
Diabetes Genetic Score		0.70 ** (0.24)	0.69 ** (0.24)	0.75 ** (0.24)	0.74 ** (0.24)
Age					
30-44		1.12 *** (0.17)	1.24 *** (0.16)	1.23 *** (0.17)	1.22 *** (0.17)
45-59		1.47 *** (0.23)	1.62 *** (0.21)	1.65 *** (0.22)	1.64 *** (0.22)
60+		-0.33 (0.38)	-0.21 (0.36)	-0.18 (0.38)	-0.20 (0.38)
BMI					
Overweight		1.24 *** (0.36)	1.21 *** (0.36)	7.66 *** (2.25)	7.64 *** (2.25)
Obese		1.90 *** (0.34)	1.87 *** (0.34)	4.77 * (2.23)	4.65 * (2.23)
Morbidly Obese		1.51 *** (0.40)	1.43 *** (0.39)	6.91 ** (2.34)	6.64 ** (2.34)
Super Morbidly Obese		2.32 ** (0.70)	2.24 ** (0.69)	14.91 *** (3.32)	14.89 *** (3.34)
Interaction Term - Glucose x BMI					
Glucose x Overweight				-0.05 ** (0.02)	-0.05 ** (0.02)
Glucose x Obese				-0.02 (0.02)	-0.02 (0.02)
Glucose x Morbidly Obese				-0.04 * (0.02)	-0.04 * (0.02)
Glucose x Super Morbidly Obese				-0.09 *** (0.02)	-0.09 *** (0.02)
AIC	1766.17	1228.70	1227.29	1195.88	1189.27
AUC	0.50	0.87	0.86	0.87	0.87
Deviance	1764.2	1196.7	1205.3	1165.9	1159.3
Log Like	-882.09	-598.35	-602.65	-582.94	-579.64
N	1383	1383	1383	1383	1382

Figures in parenthesis indicate standard errors
Significance Codes: *** p<0.001; ** p<0.01; * p<0.05

BMI = Body Mass Index

Outlier Identification

High leverage points were investigated relative to the best performing model (model 4). Patients 2333 and 2356 displayed both high leverages and large residuals in the leverage versus residuals plot. However, all features lay within biologically plausible ranges and thus the observations were not removed.

Cook's distance was used to measure the overall influence of observations on the regression model. Patient 2360 was identified as moderately influential with a Cook's distance of 0.05, due to a biologically implausible triceps skinfold thickness of 110mm. DFBETAs for skin thickness were calculated to assess the effect of the data points on the coefficient. Although many observations lay above the calculated cut-off of 0.05, patient 2360 stood out significantly (absDFBETAS=0.45). Patient 2360 was removed as an outlier and the final model re-fitted, model 5 (see Table 1.2).

Model Interpretation

Terms with odds ratio (OR) 95% confidence intervals (CIs) crossing 1 were deemed statistically insignificant.

Respectively individuals aged “30-44” (OR=3.39, 95%CI [2.45, 4.70]) and “45-59” have approximately 3.4 times and 5.2 times the risk of having diabetes compared to those the reference category of “18-29”. For individuals aged “60+” (OR=0.82, 95%CI [0.38, 1.68]) the risk is reduced by 18%, though this result is not statistically significant.

Genetic predisposition is a strong predictor of diabetes. A unit increase in diabetes genetic score (OR=2.09, 95%CI [1.31, 3.36]) is associated with a 109% increased risk in having diabetes. A unit (mm) increase in triceps skinfold thickness (OR=1.05, 95%CI [1.03, 1.07]) is associated with a 5% increased risk.

The inclusion of the interaction term means that the OR of glucose (OR=1.07, 95%CI [1.04,1.11]) represents it’s effects when BMI is at the reference category, “Normal Weight”. Within this range a unit (1 mg/dL) increase in glucose is associated with a 7% higher diabetes risk. As a glucose value of 0 is implausible, the ORs of the BMI categories cannot be interpretated literally. However, the general trend indicates that increasing BMI is associated with an increasing risk of diabetes.

To understand the relationship between glucose and BMI, an interaction plot was created (Figure 1.1). For all BMI categories except 'Super Morbidly Obese,' an increase in glucose raises diabetes probability, whereas in the 'Super Morbidly Obese' category, it decreases. The interaction terms for all BMI groups, except the “Obese” category, are significant.

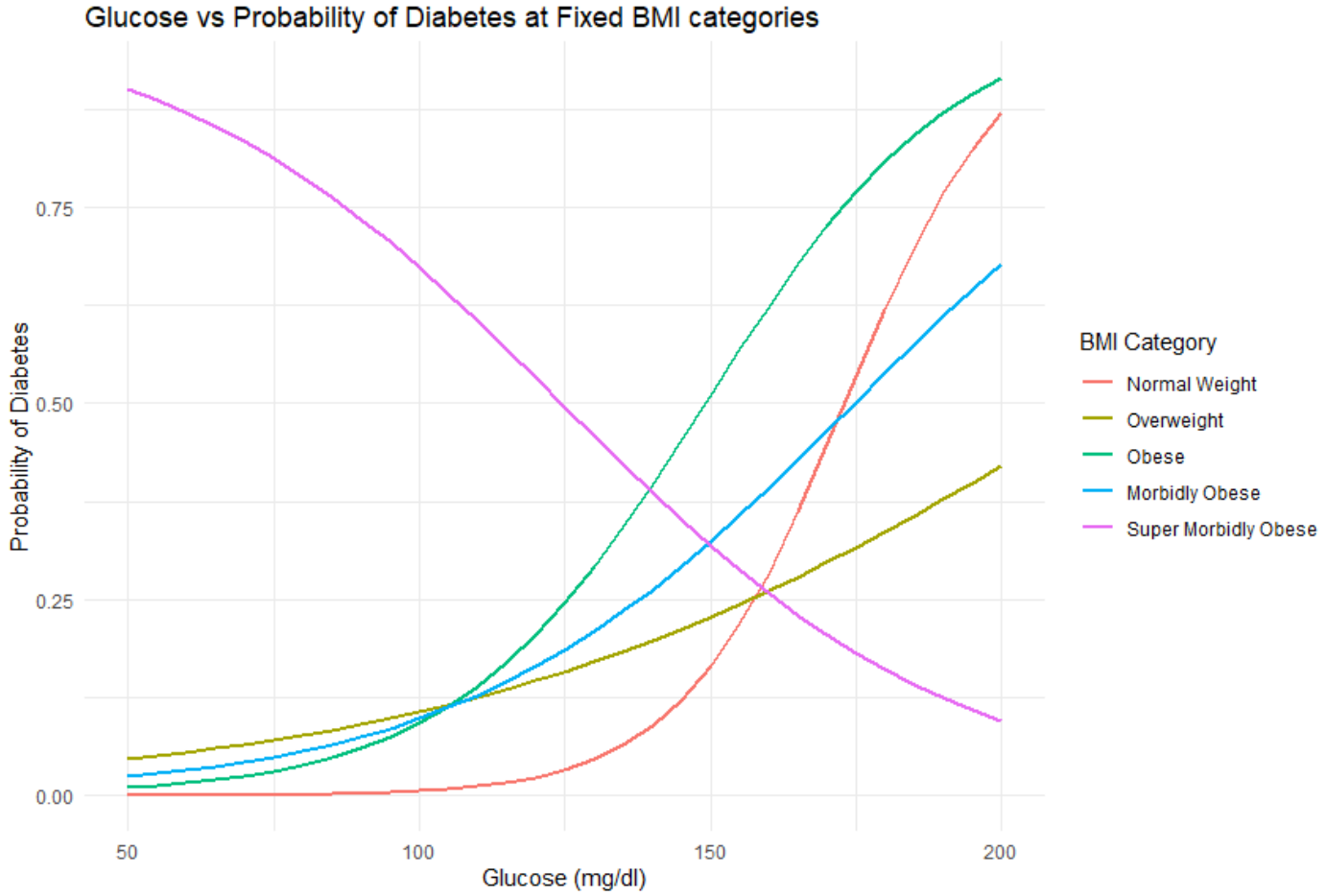

Figure 1.1. Interaction between glucose and BMI. Age and number of pregnancies are fixed at their respective reference categories, “18-29” and “0”. Triceps skinfold thickness and genetic diabetes score are fixed at mean values, 29.18 mm and 0.47 respectively.

Model Discrimination and Calibration

Model discrimination was assessed using an ROC curve. AUC values of 0.88 (train) and 0.86 (test) show good discrimination.

At a 0.5 threshold, sensitivity dropped from 60.0% (train) to 55.8% (test), while specificity remained high at 89.5% (train) and 89.6% (test). The model minimises false positives but misses approximately 40% of actual positive cases. To address this, a range of threshold values were compared. A threshold of 0.3 provided balanced results for both the train (sensitivity=82.7%, specificity=76.3%) and test set (sensitivity=76.9%, specificity=79.8%).

The high AUCs do not appear to align with the sensitivity analysis results. The AUC remains high because the model performs very well for the majority class, no diabetes, as shown by high specificity.

The Hosmer-Lemeshow test was performed to assess model calibration. For the train set (p=0.4) one fails to reject the null hypothesis that there is no significant difference between the observed and expected outcomes, indicating good calibration for the train dataset.

However, a significant result (p=4.5e^{-5}) indicates that the model is not very generalisable. A significant contributor to poor generalisability is likely to be the high data redundancy.

The Hosmer-Lemeshow test does not specify the direction of miscalibration. The train calibration plot is symmetric about the mean (calibration-in-the-large=0), while the test plot (calibration-in-the-large=-0.03) shows diabetes probability is generally underpredicted.

Model Robustness

Sensitivity analysis was performed for BMI and age categories to assess model robustness at sub-group level. The "Overweight" category has very poor sensitivity, at 28.6% (train) and 19.7% (test). In the test set, sensitivity dropped sharply for the "Normal Weight" category (78.5% to 37.5%), while specificity fell for the "Super Morbidly Obese" group (71.4% to 33.3%).

The training data has poor sensitivity for both "18-29" (train=37.1%, test=35.1%) and "60+" (train=55.6%, test=31.3%) age groups.

A sharp drop in sensitivity is observed for marginal groups; "Super Morbidly Obese" and age "60+", which were under-represented within the dataset. Poor generalisability is also observed for both reference categories; age "18-29" and "Normal Weight". These groups likely have a lower prevalence of the outcome, and a lack of strong predictive features, which can lead to misclassification.

Conclusion

Analysis of the final model reveals that genetic diabetes score, and blood glucose concentration are strong predictors of diabetes. A glucose x BMI interaction term showed that an increase in glucose is associated with a decrease in diabetes probability for only one BMI category, "Super Morbidly Obese." High AUC values for both train and test datasets indicate good discrimination. However, failure of the Hosmer-Lemeshow test for the test data and poor sensitivity, highlight poor model calibration and generalisability. It is postulated that excessive redundancy in the dataset prevented the model from identifying true trends, instead overfitting to repetitive patterns.

Appendix

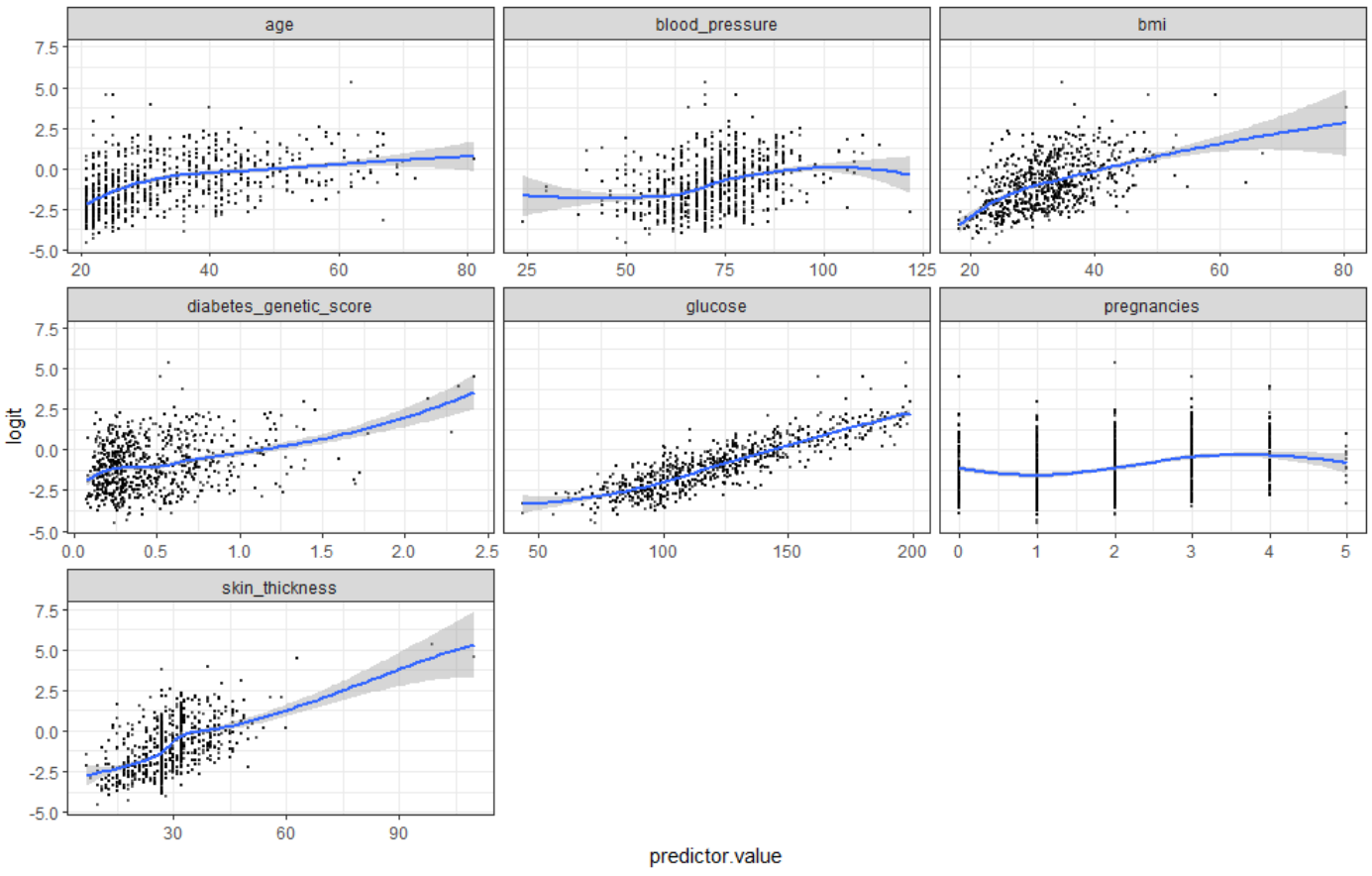

Figure 2.1. Predicted log-odds versus continuous variables plots

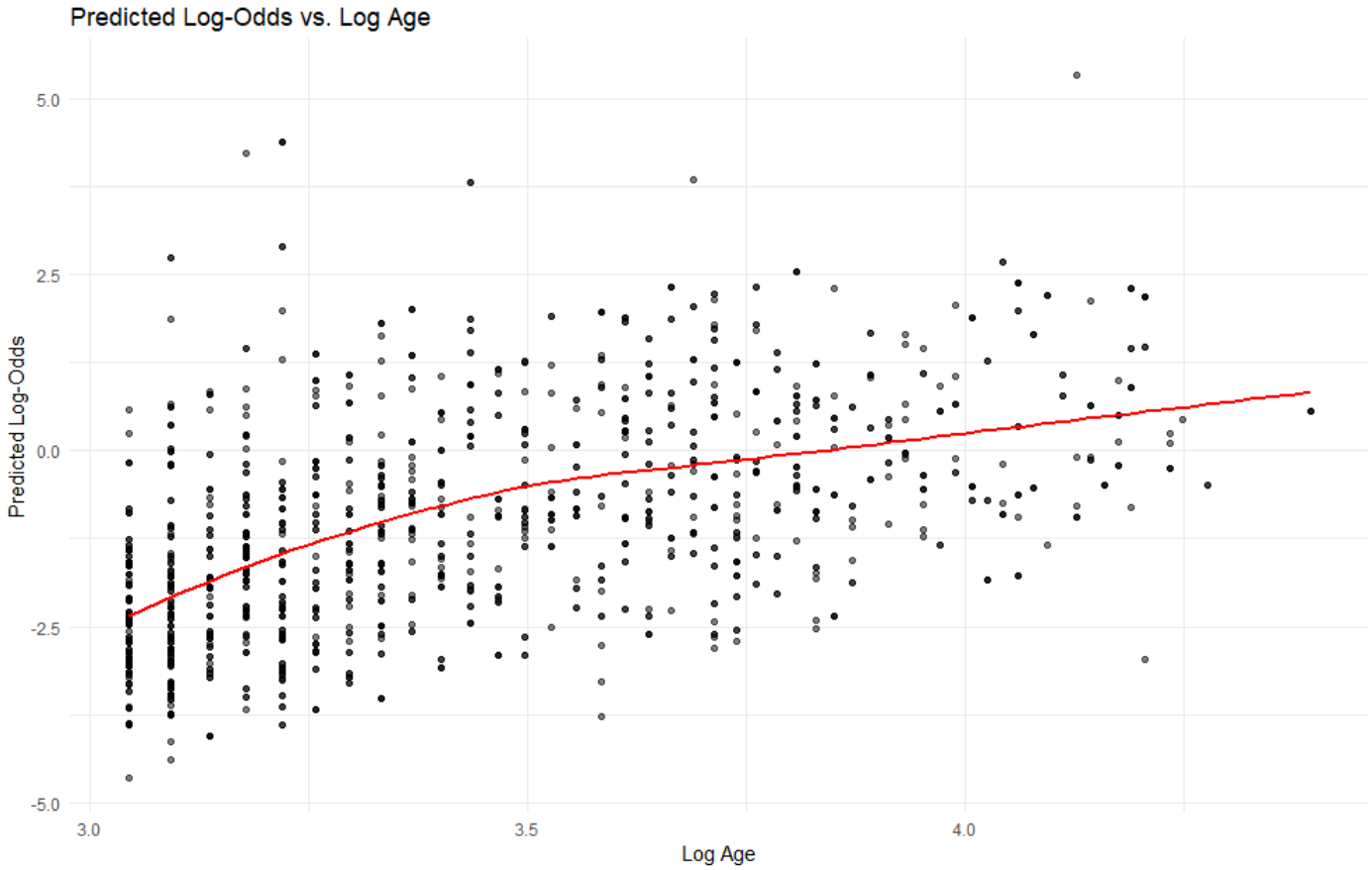

Figure 2.2. Plot of predicted log-odds versus log-age