Diabetes Risk Prediction Using Discriminant Analysis

Dishari Datta, Iman Kalyan Dutta, Tanmoy Nath

April 24, 2025

Introduction: Background

- **Global Disease Burden**:
  - Diabetes prevalence has nearly tripled since 2000
  - Rapid increase especially in low- and middle-income countries
- **Economic and Health Impacts**:
  - $966 billion global healthcare cost in 2021 (12% of spending)
  - 2–4x higher risk of cardiovascular disease
  - Up to 10 years shorter lifespan (type 1 diabetes)
- **Diagnostic Challenges**:
  - 44% of diabetes cases remain undiagnosed
  - Testing requires clinical visits (fasting glucose, HbA1c)
  - Delayed diagnosis causes irreversible complications

Data Collection and Pre-processing

Data Source: NHANES

- Database: National Health and Nutrition Examination Survey (NHANES) 2013–2017
- Accessed via R package: `nhanesA`

Example R Code:

```
nhanes("GHB_J")
nhanesCodebook("GLU_J")
```

- **Domains Included in NHANES Database:**
  - Demographics: Age, gender, ethnicity
  - Examination data: BP, body measurements
  - Laboratory: Blood biomarkers
  - Dietary intake: Nutrient consumption
  - Questionnaire: Lifestyle & medical history

Feature Engineering

- 33 covariates selected from NHANES database.
- Grouped into the following categories:

Category	**Features**
Demographics	age, gender, ethnicity, `income_poverty_ratio`, `household_size`
Physical Activity	`sedentary_minutes`, `activity`, `recreational_activity`
Dietary Intake	`calorie`, `carbs`, `sugar`, `fibre`, `saturated_fat`, `trans_fat`, `protein`
Biochemical Markers	`trigly`, `ldl_chol`, `total_chol`, `hdl_chol`, `creatinine`, `albumin`
Anthropometrics	`bmi`, `waist`
Blood Pressure	`systolic_bp`, `diastolic_bp`
Lifestyle Factors	`alcohol_status`, `family_history`

Table: Feature categories used for modeling

* `activity` and `recreational_activity` are derived features.

Derived Feature: `activity`

- Summarizes day-to-day physical activity.
- Based on:
  - `PAQ605` – Moderate activity (e.g., brisk walking, gardening)
  - `PAQ620` – Vigorous activity (e.g.,construction worker,labor intense job)
- Scoring logic:
  - 2 = High activity: PAQ605 = "No", PAQ620 = "Yes"
  - 1 = Moderate activity: PAQ605 = "Yes", PAQ620 = "No"
  - 0 = No activity: Both = "No"
  - NA = Missing or undefined

Derived Feature: `recreational_activity`

- Captures recreational physical activity.
- Based on:
  - `PAQ650` – Moderate recreational (e.g., biking slowly, dancing)
  - `PAQ665` – Vigorous recreational (e.g., running, tennis, weight-lifting)
- Scoring logic:
  - 2 = High: PAQ650 = "No", PAQ665 = "Yes"
  - 1 = Moderate: PAQ650 = "Yes", PAQ665 = "No"
  - 0 = None: Both = "No"
  - NA = Missing or undefined

Response Variable Construction

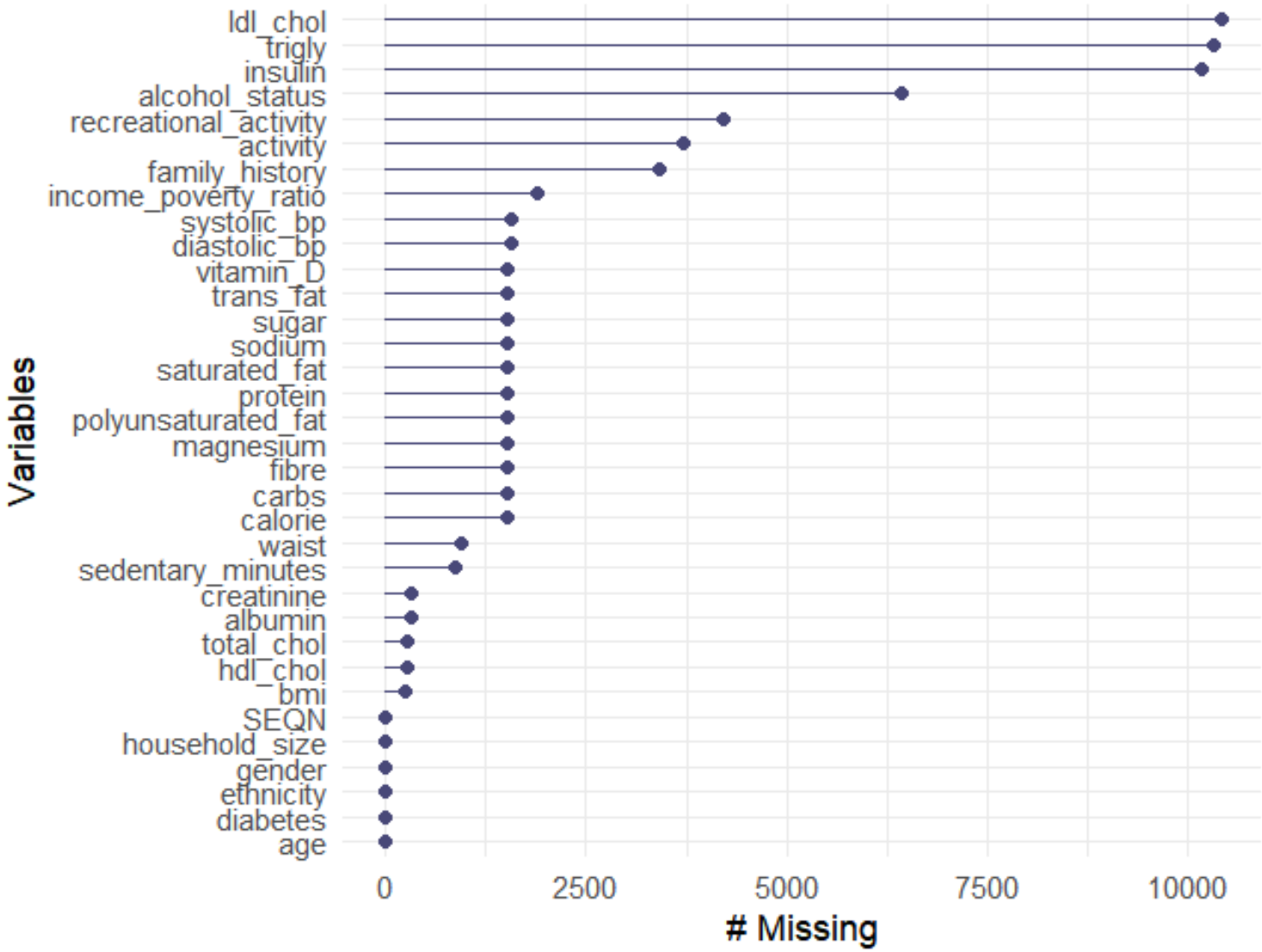
- Diabetes classification based on:
  - **HbA1c** (Glycohemoglobin) from GHB_*
  - **Fasting Plasma Glucose (FPG)** from GLU_*
- Classification criteria:
  - **Diabetic (2)**: HbA1c $> 6.4\%$ or Glucose > 125 mg/dL
  - **Prediabetic (1)**: HbA1c 5.7–6.4% and/or Glucose 100–125 mg/dL
  - **Normal (0)**: HbA1c $< 5.7\%$ and Glucose < 100 mg/dL
- Missing data handling:
  - If one value missing $\rightarrow$ classify with available measure
  - If both missing $\rightarrow$ exclude from analysis

Classification Logic in R

```
diabetes = case_when(
  LBXGH > 6.4 | LBXGLU > 125 ~ 2,  # Diabetes
  (LBXGH < 5.7 & LBXGLU >= 100 & LBXGLU <= 125) |
  (LBXGH >= 5.7 & LBXGH <= 6.4 & LBXGLU < 100) |
  (LBXGH >= 5.7 & LBXGH <= 6.4 &
   LBXGLU >= 100 & LBXGLU <= 125) ~ 1,  # Prediabetes
  (LBXGH < 5.7 & LBXGLU < 100) ~ 0  # Normal
)
```

Missing Data Overview

- Visualized missingness using `naniar:: gg_miss_var()` in R



Imputation Strategy

- **Continuous variables**: Median imputation
- **Categorical variables**: Mode imputation
- Context-aware sub-grouping based on:
  - `age_group`, `gender`, `bmi_category` etc.
  - Condition-specific risk categories (e.g., `kidney_risk`)
- Goal: Preserve structure and reduce bias

Imputation Details by Variable (1/2)

- **income_poverty_ratio (Median)**: Grouped by `household_size`, `ethnicity`, `age_group`
- **sedentary_minutes (Median)**: Grouped by `age_group`, `bmi_category`, `activity`
- **activity (Mode)**: Grouped by `age_group`, `gender`, `bmi_category`
- **recreational_activity (Mode)**: Grouped by `age_group`, `bmi_category`
- **Diet variables (Median)**: Grouped by `age_group`, `bmi_category`, `diabetes`
- **alcohol_status (Mode)**: Grouped by `age_group`, `gender`, `diabetes`

Imputation Details by Variable (2/2)

- **total_chol, hdl_chol (Median)**: Grouped by `age_group`, `gender`, `bmi_category`
- **trigly, ldl_chol (Median)**: Based on lipid group & `diabetes`
  Lipid group defined as:
  - Normal: `total_chol` < 200 and `hdl_chol` >= 40
  - High-risk: `total_chol` >= 200 or `hdl_chol` < 40
- **bp (Median)**: Grouped by `age_decade`, `gender`, `bp_risk_group`
- **creatinine, albumin (Median)**: Grouped by `gender`, `age_group`, `kidney_risk`
  Kidney risk levels: Diabetic, Hypertensive (SBP $\geq$ 140), Low-risk
- **bmi, waist (Median)**: Grouped by `bodycomp_group`
  Defined as: Diabetic, Older_Male (age $\geq$ 50), Older_Female (age $\geq$ 50), General
- **family_history (Mode)**: Grouped by `age_group`, `ethnicity`

Exploratory Data Analysis

Exploratory Data Analysis: Class Distribution

- The dataset shows an imbalanced distribution across diabetes categories:

Category	Percentage
Normal	52.3%
Prediabetic	31.7%
Diabetic	16.0%

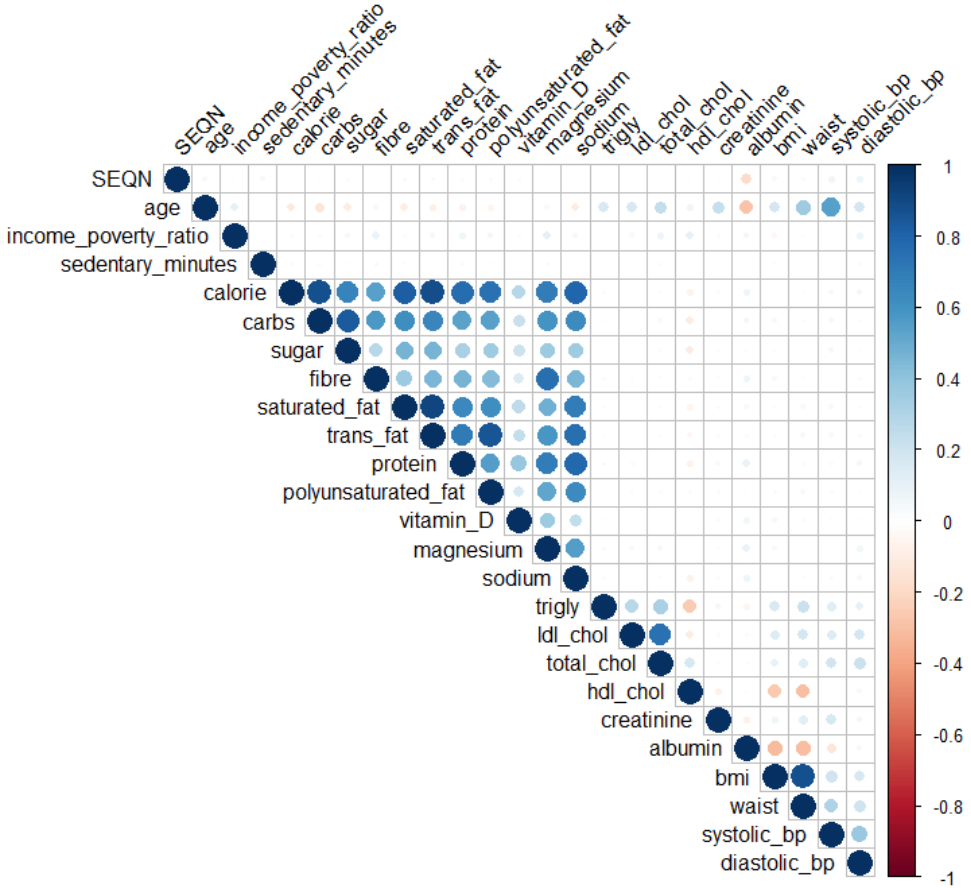
Table: Class distribution of the response variable

Correlation Analysis

- Explore interdependencies among numeric features
- Identify:
  - Redundant variables
  - Detect Multicollinearity
- Important for parametric models like LDA and QDA which assume feature independence

Correlation Matrix

- Computed correlations among all numeric variables



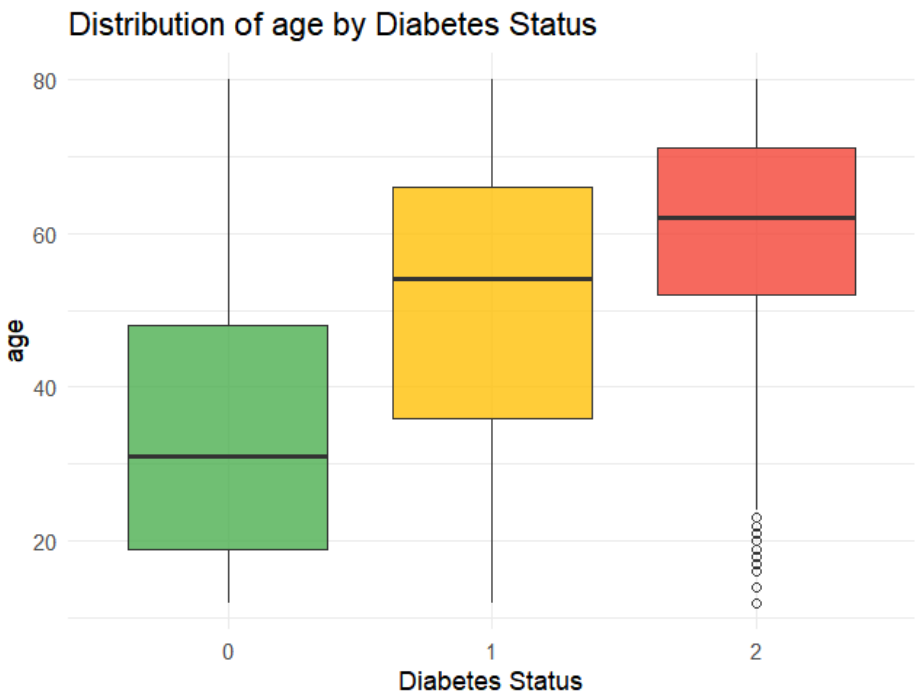
Key Observations and Implications

Notable Correlations:

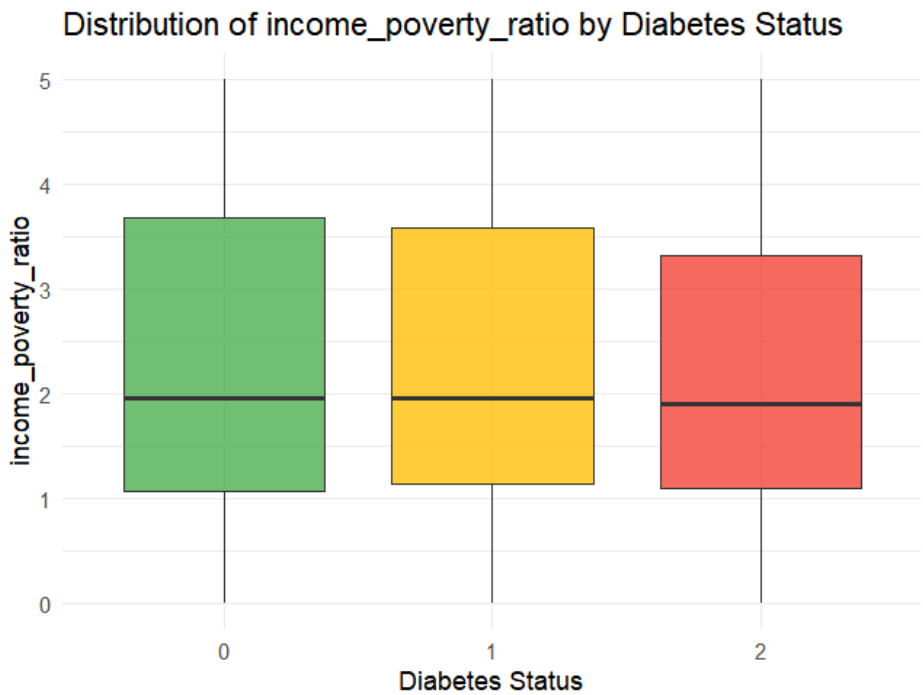
- **Blood Pressure:** Systolic & diastolic BP are strongly correlated
- **Anthropometry:** BMI and waist circumference highly correlated
- **Diet:** Calories, carbs, fats form a coherent intake cluster
- **Lipid Profile:** Total chol. & LDL strongly correlated
- **Orthogonal Features:** sedentary minutes and income ratio show weak correlation

Linearity Assessment: Demographic Features

- Boxplots used to examine distribution of continuous variables by diabetes category

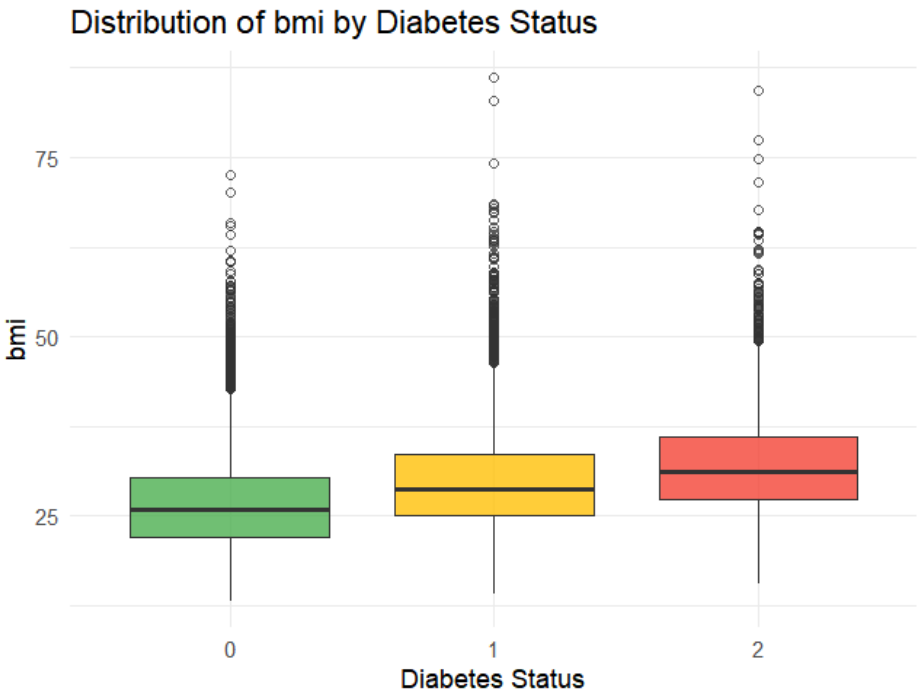


Age: Strong positive trend

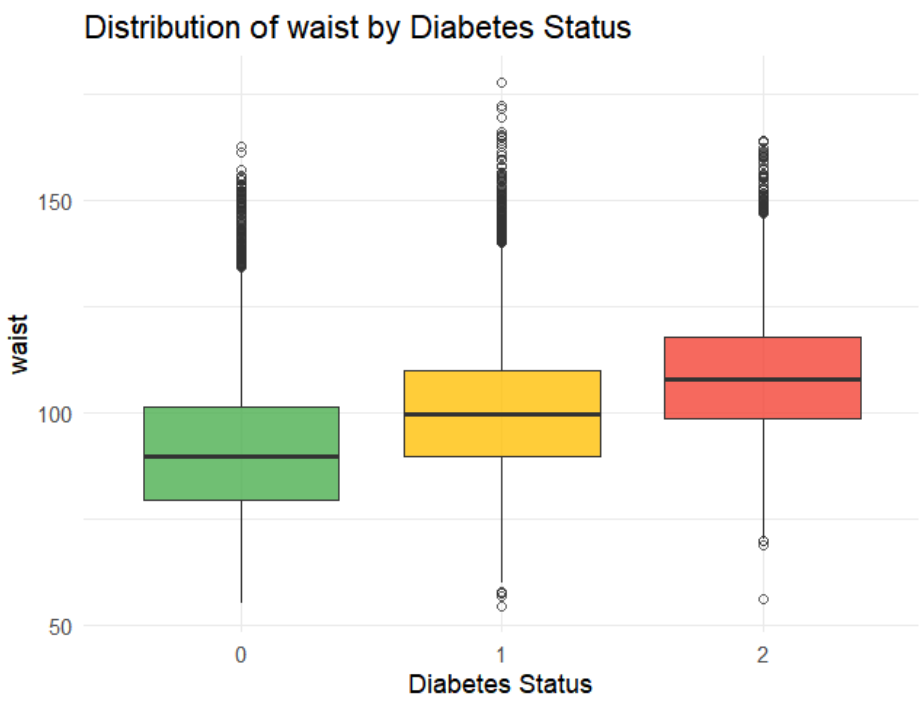


Income-Poverty Ratio: Inversely related

Linearity Assessment: Anthropometric Measures

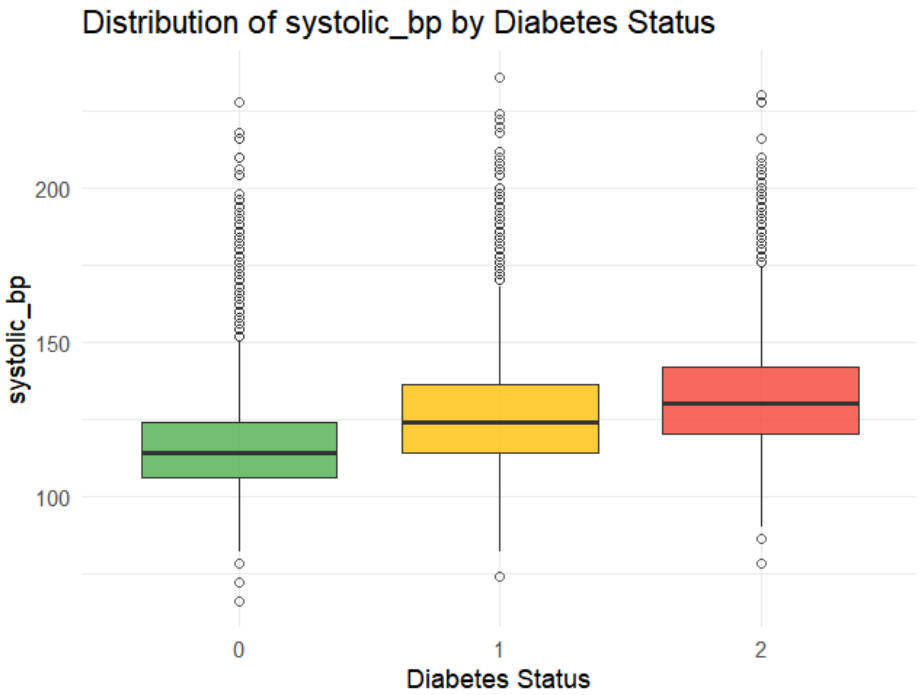


BMI: Progressive increase with severity

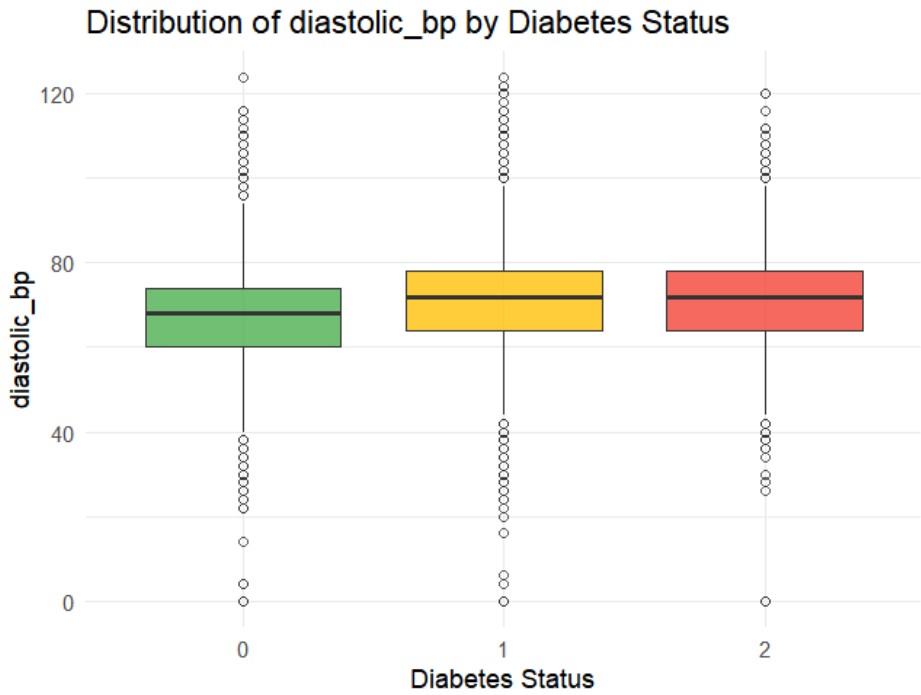


Waist: Mirrors BMI trend

Linearity Assessment: Blood Pressure

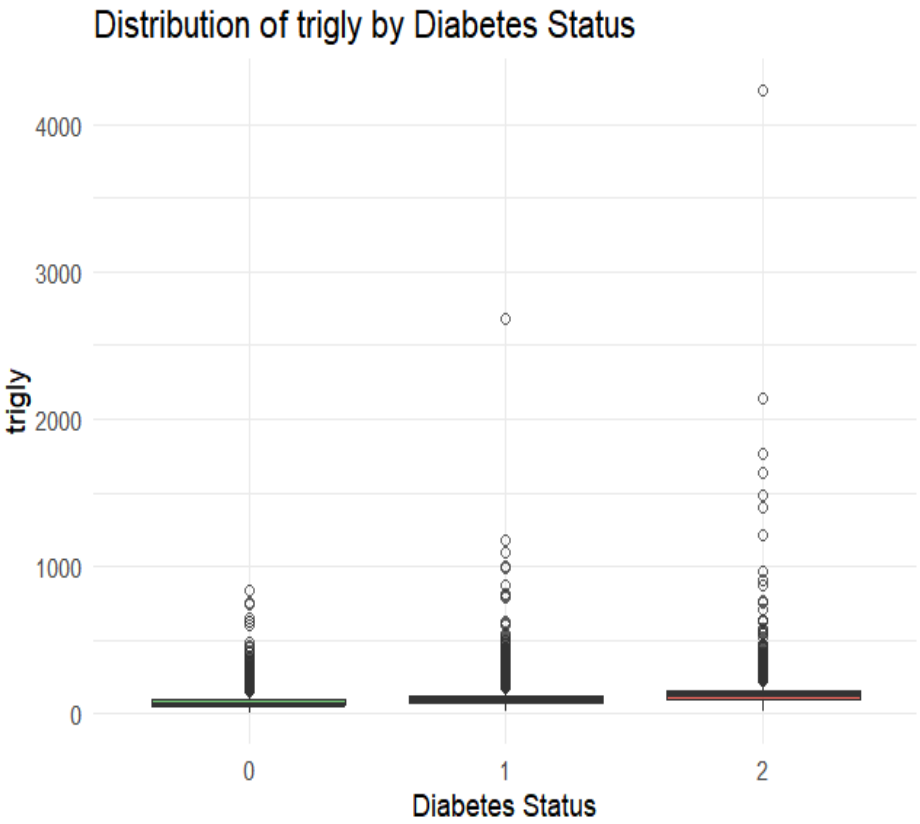


Systolic BP: Moderate linear trend

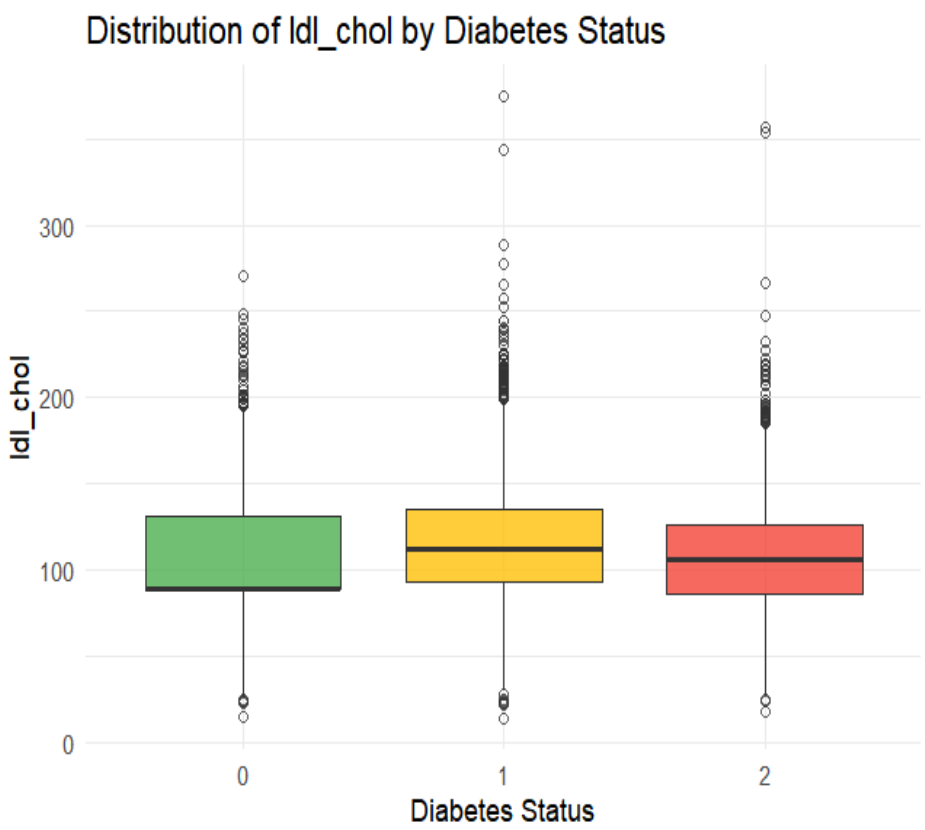


Diastolic BP: Weaker linearity

Linearity Assessment: Lipid Profiles (1/2)



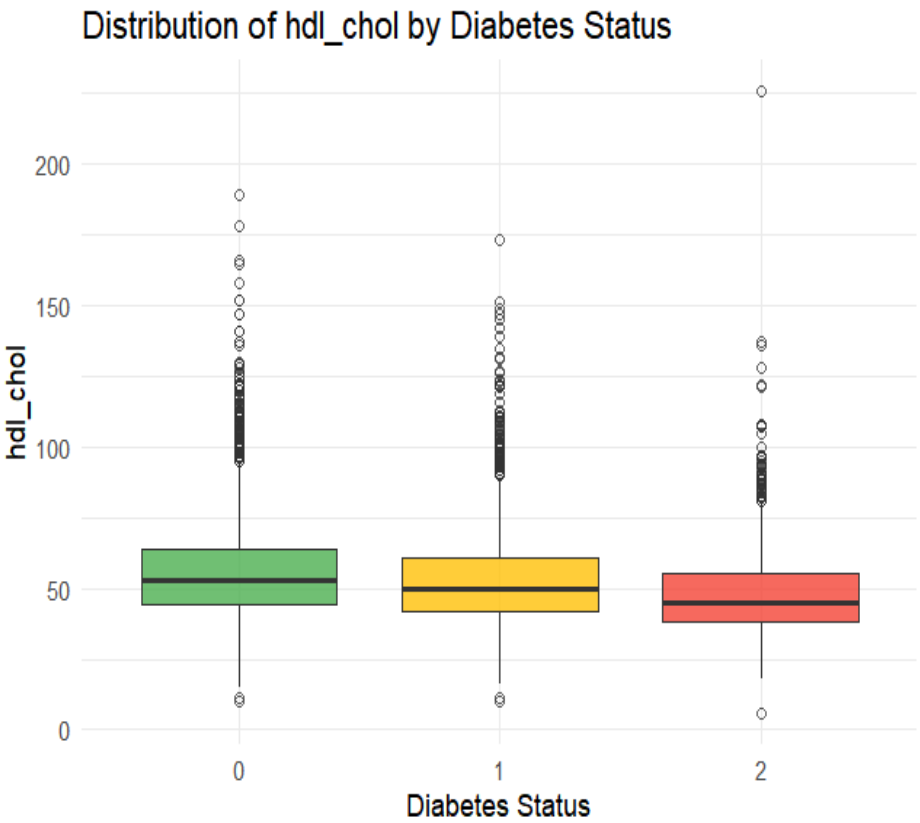
Triglycerides



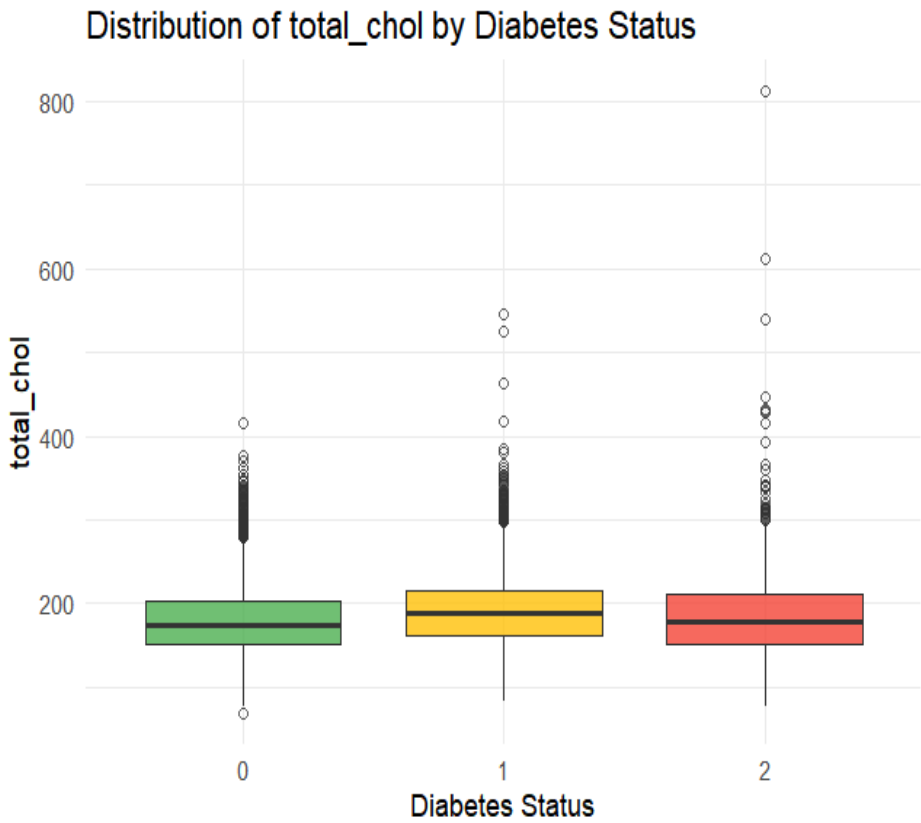
LDL Cholesterol

- **Triglycerides** show a strong positive relationship with diabetes status.
- **LDL Cholesterol** exhibits moderate association

Linearity Assessment: Lipid Profiles (2/2)



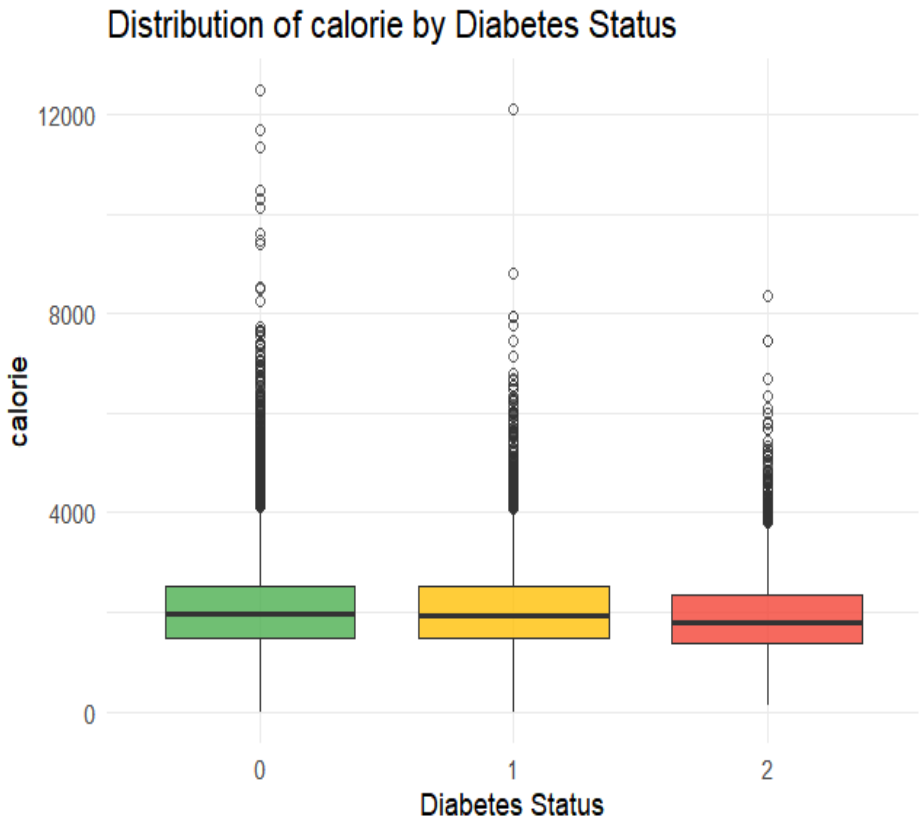
HDL Cholesterol

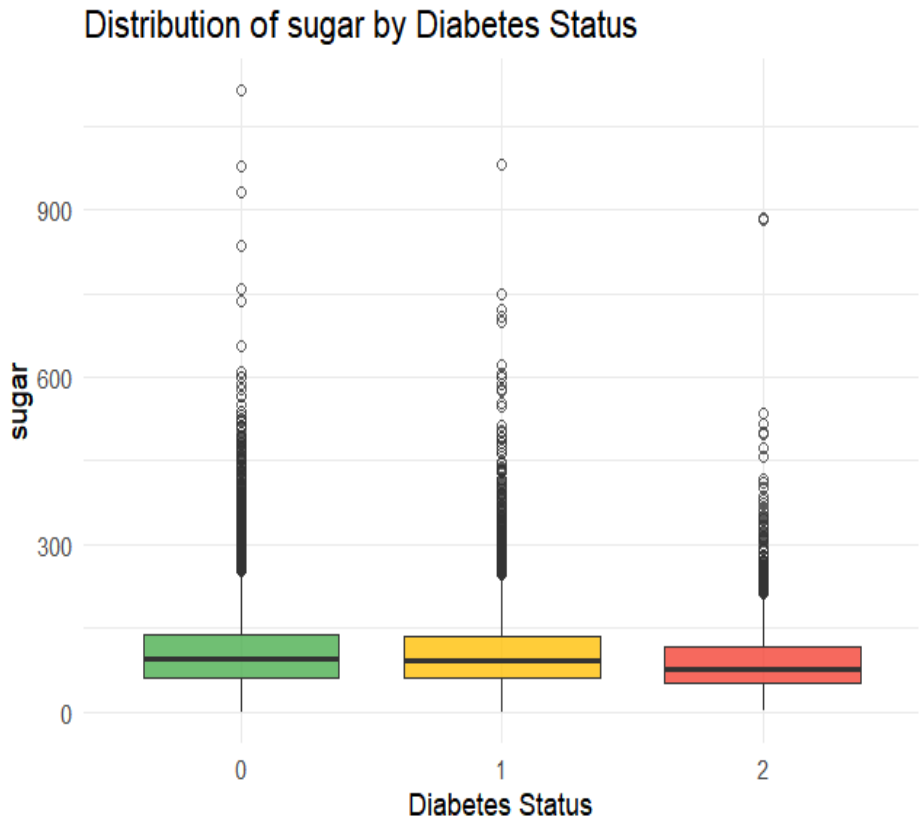


Total Cholesterol

- **HDL Cholesterol** demonstrates strong inverse relationship
- **Total Cholesterol** shows weakest association

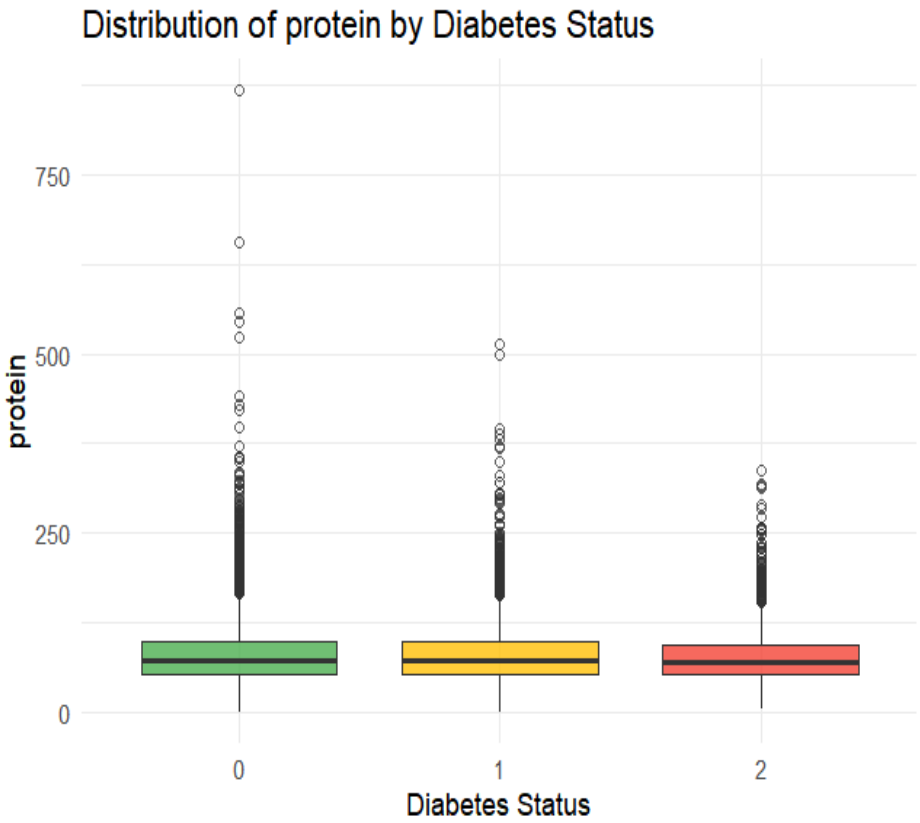
Linearity Assessment: Dietary Components (1/3)

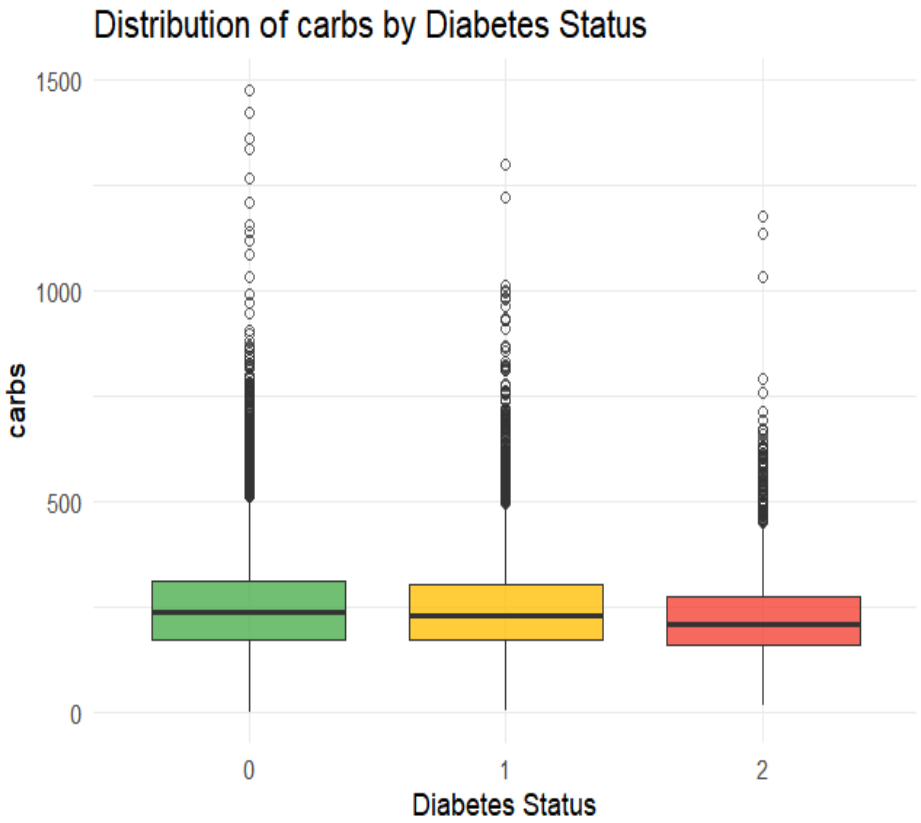




- **Calorie Intake** shows no significant differences.
- **Sugar Intake** reveals expected pattern.

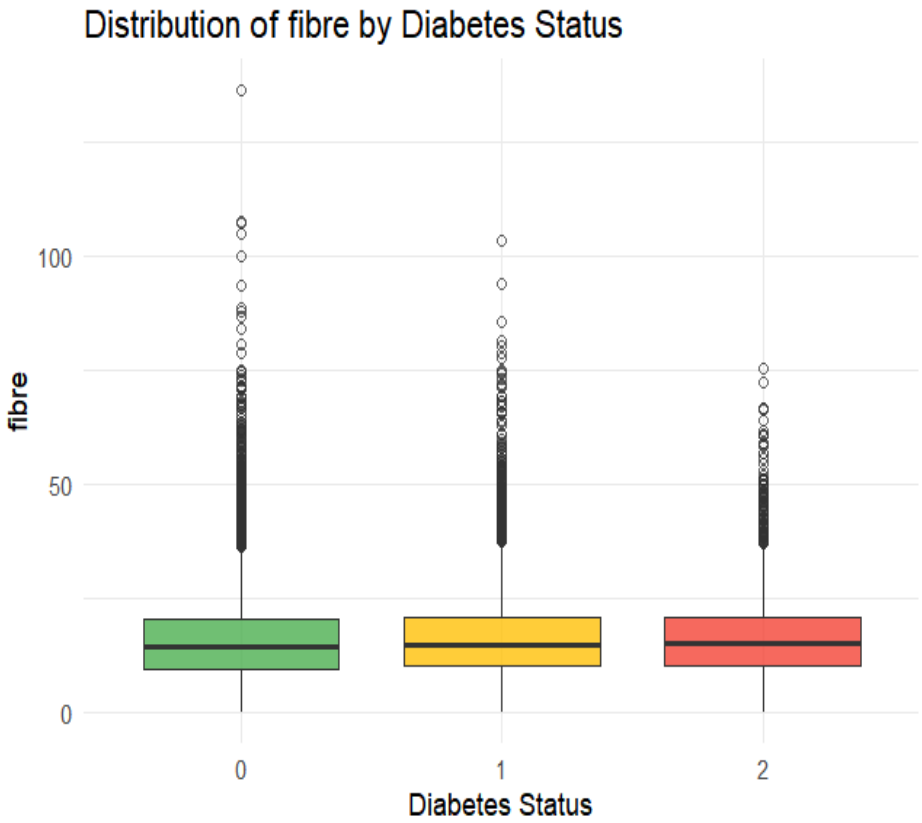
Linearity Assessment: Dietary Components (2/3)

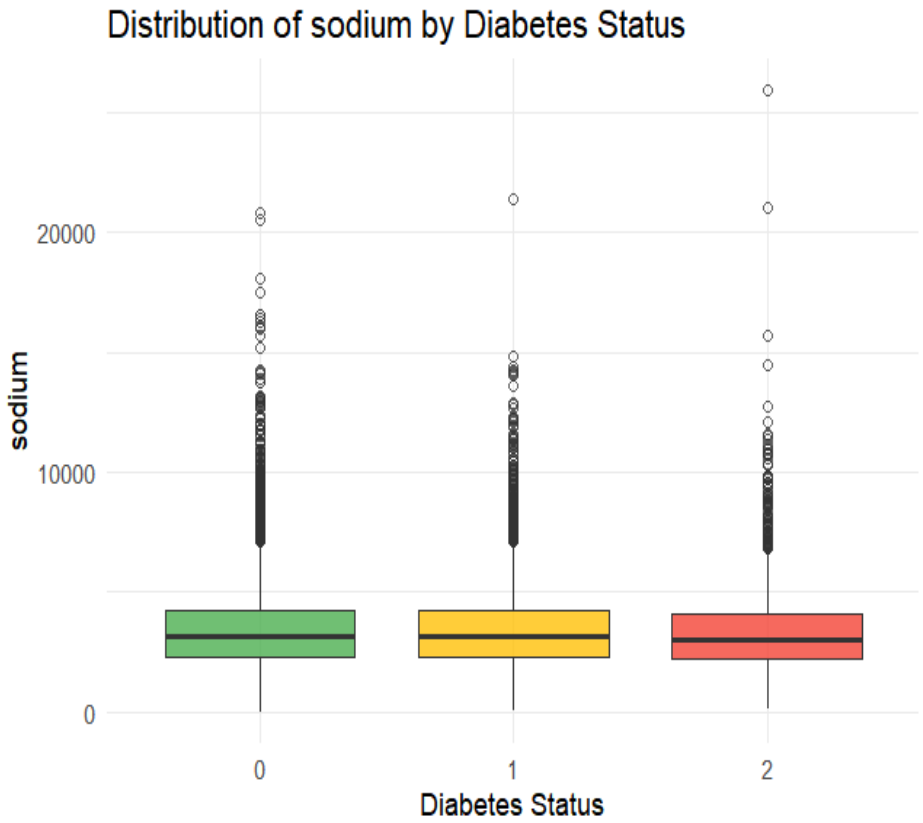




- **Protein** shows minimal variation
- **Carbohydrates** display slight positive association

Linearity Assessment: Dietary Components (3/3)

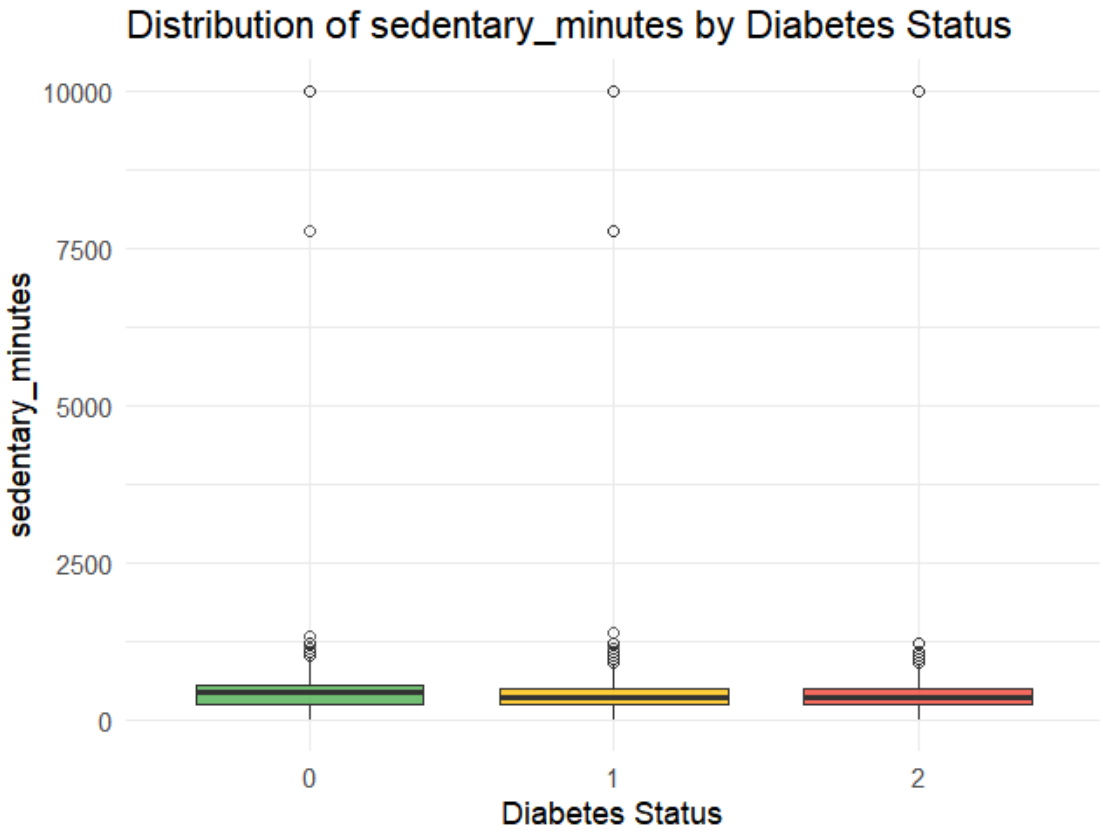




- **Dietary Fiber** demonstrates inverse relationship.
- **Sodium** exhibits no discernible pattern.

Linearity Assessment: Physical Activity

- Higher variance observed in the diabetic group



Summary of Linearity Findings

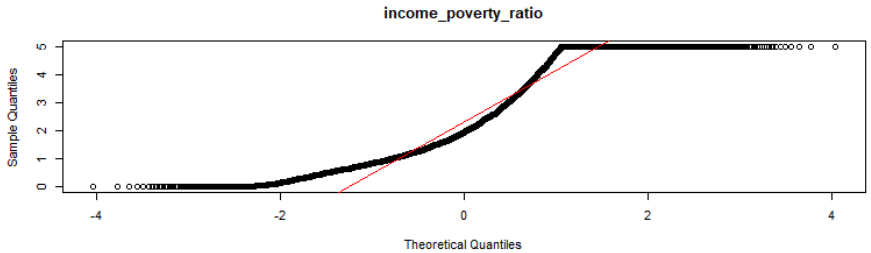
Feature Category	Key Observations
Demographics	Age shows strongest linear relationship
Anthropometrics	All measures increase with diabetes status
Blood Pressure	Systolic BP more informative than diastolic
Lipid Profiles	Triglycerides and HDL most discriminative
Dietary Factors	Generally weak predictors individually

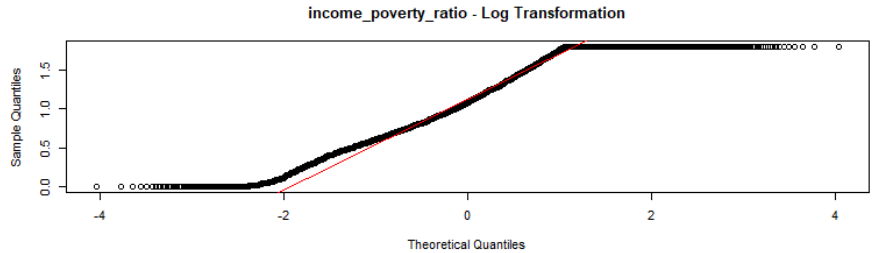
Implications for Discriminant Analysis

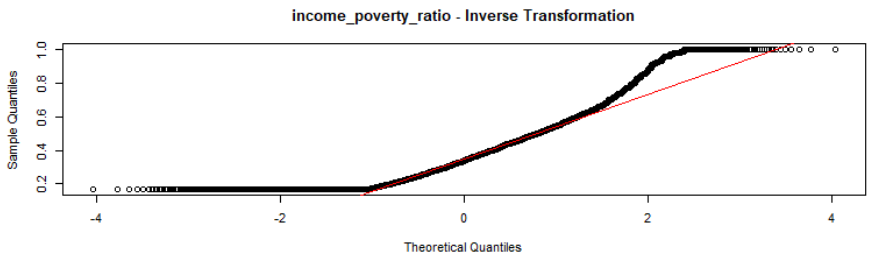
- **LDA suitability:**
  - Strong linearity in age, BMI, HDL, and triglycerides supports LDA assumptions
  - Class-wise variance in features like age may still impact performance
- **QDA considerations:**
  - Non-linear features and class-specific variances (e.g., biochemical markers)
  - Robust to outliers and distributional violations

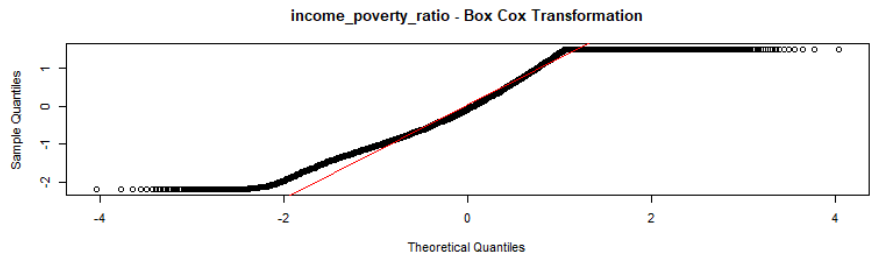
Univariate Normality: Income Poverty Ratio

- Q-Q plots analysed for five transformations:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson



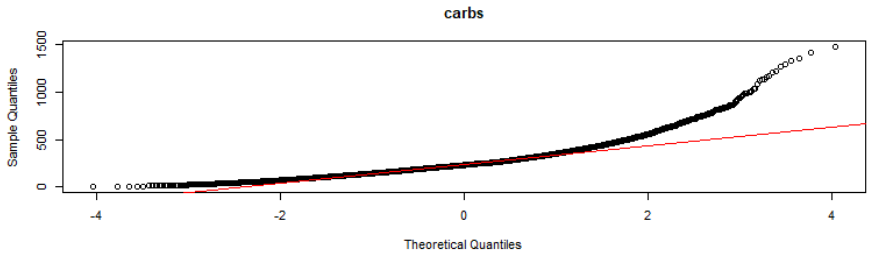


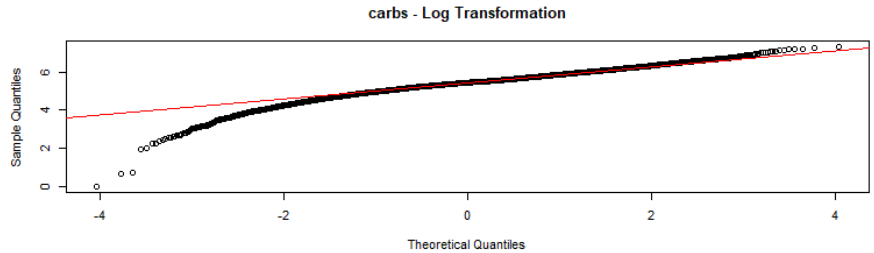


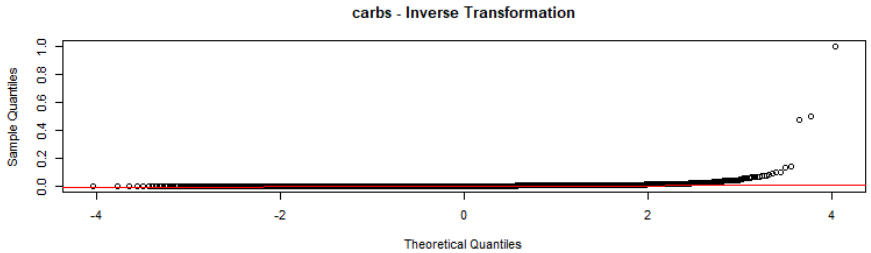


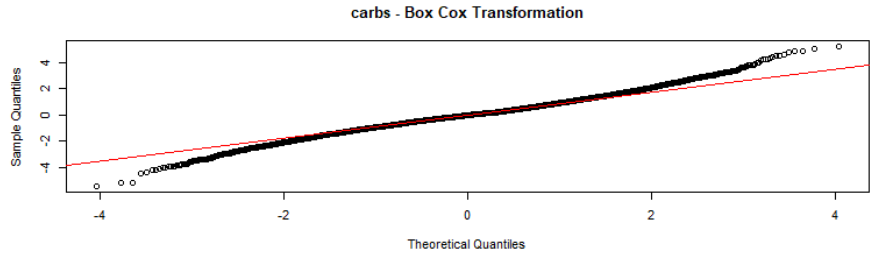
Univariate Normality: Carbohydrates

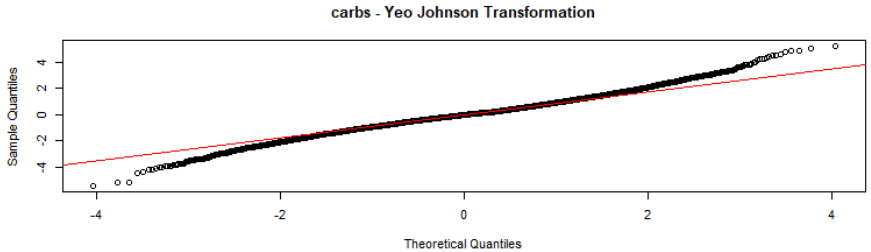
- Q-Q plots analyzed for five transformations:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson





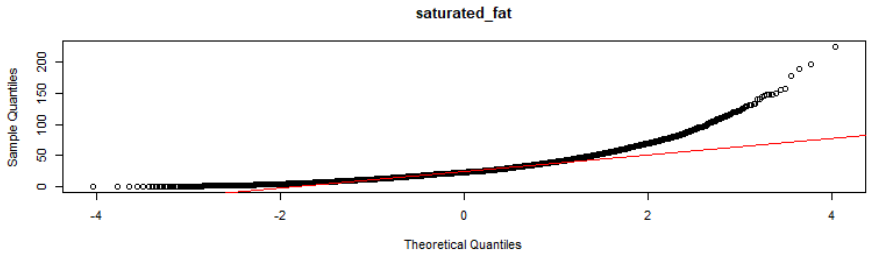


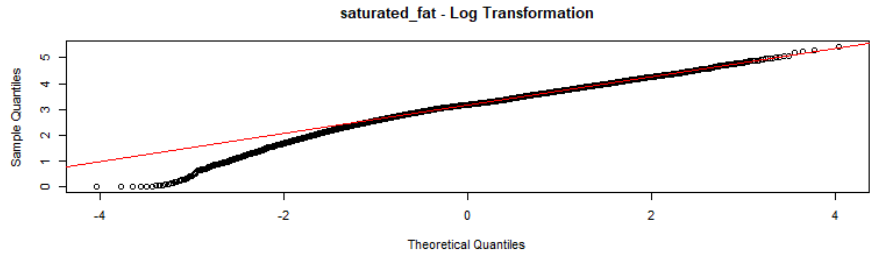


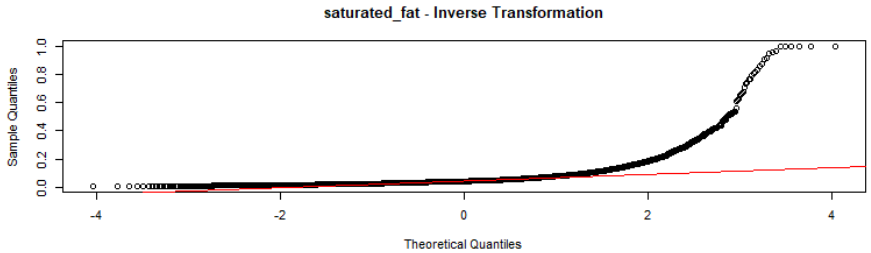


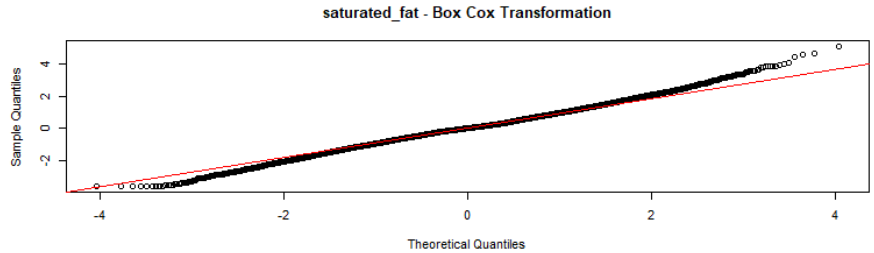
Univariate Normality: Saturated Fat

- Q-Q plots under 5 transformations:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson



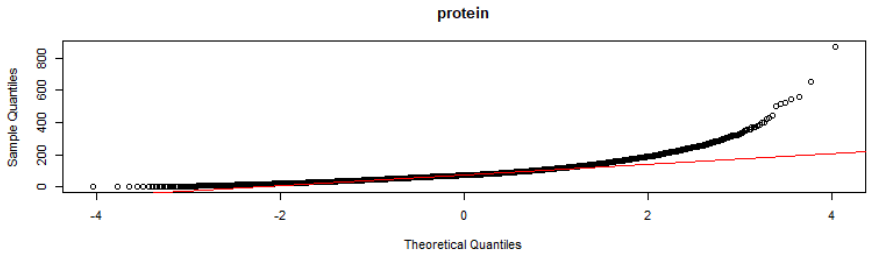


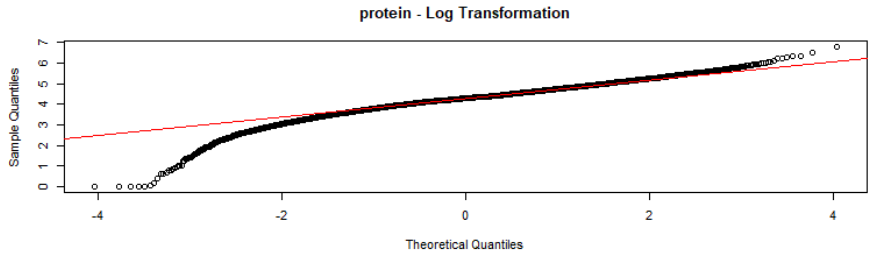


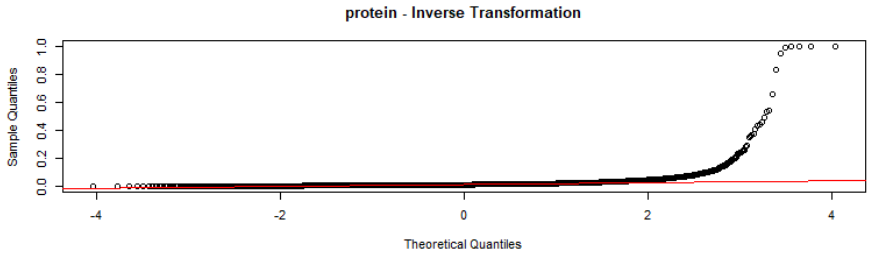


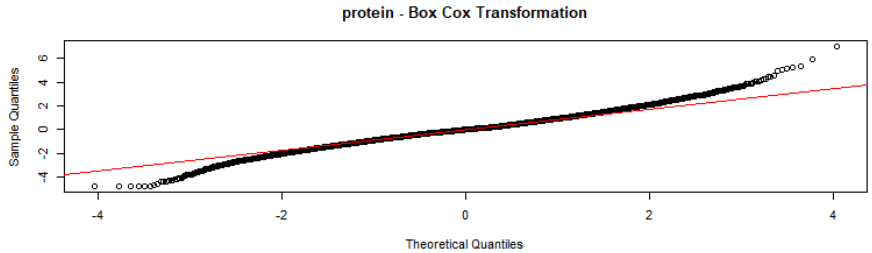
Univariate Normality: Protein

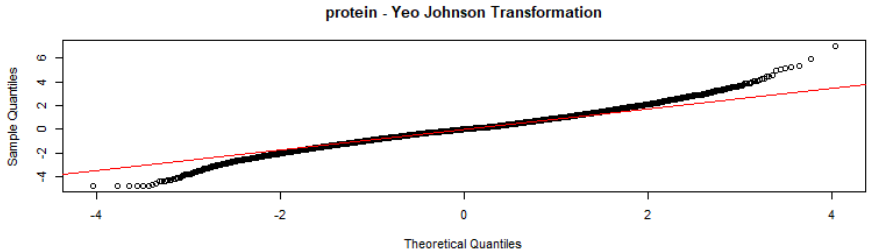
- Evaluated using Q-Q plots under:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson transformations





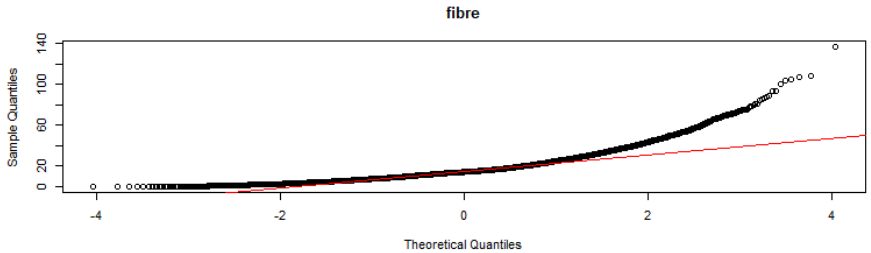


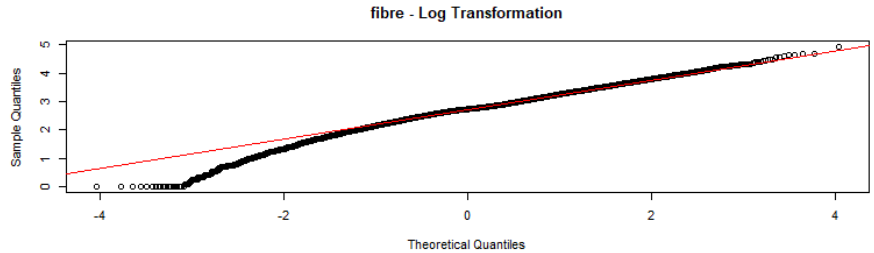


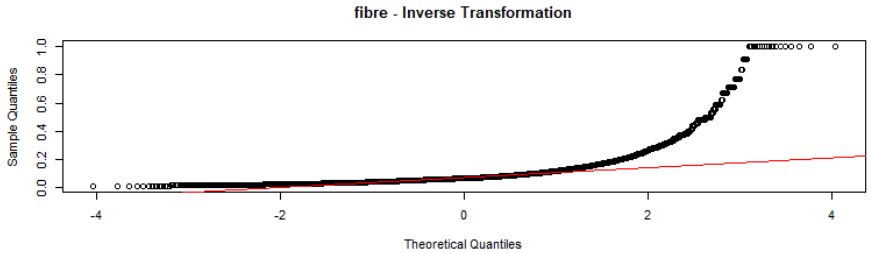


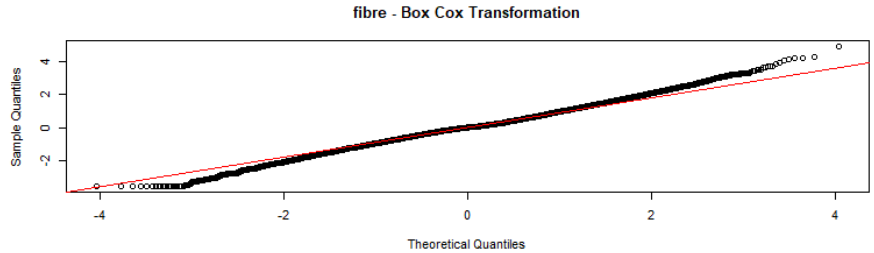
Univariate Normality: Fibre

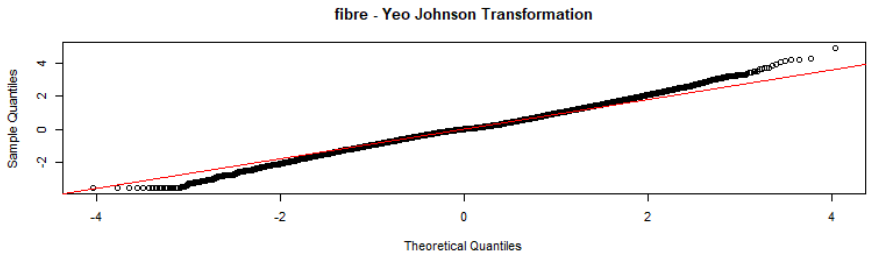
- Q-Q plots examined under five transformations:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson





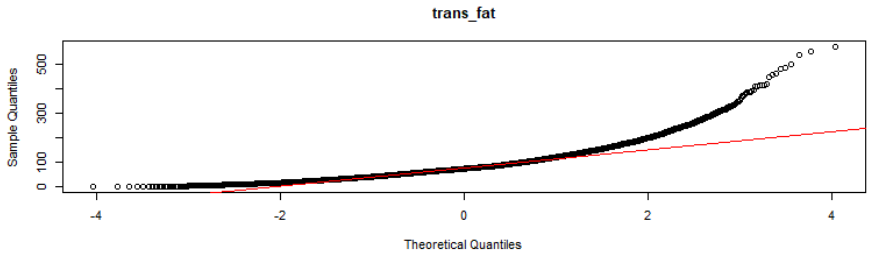


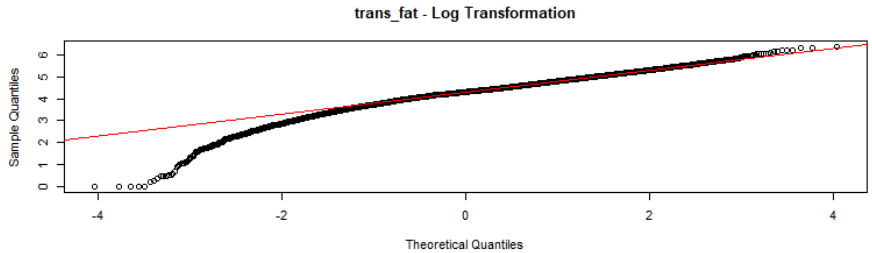


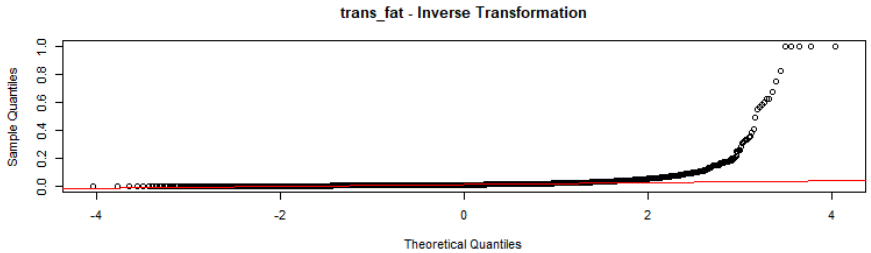


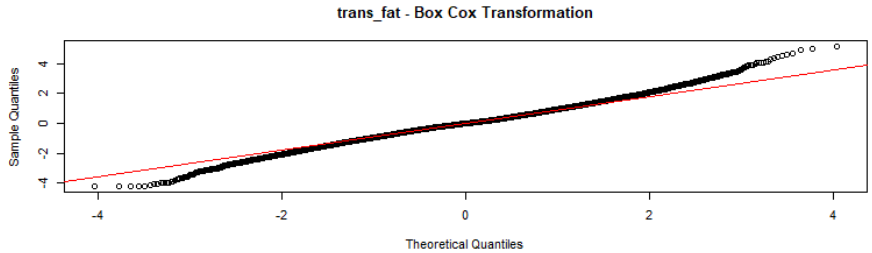
Univariate Normality: Trans Fat

- Assessed with Q-Q plots under five transformations:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson



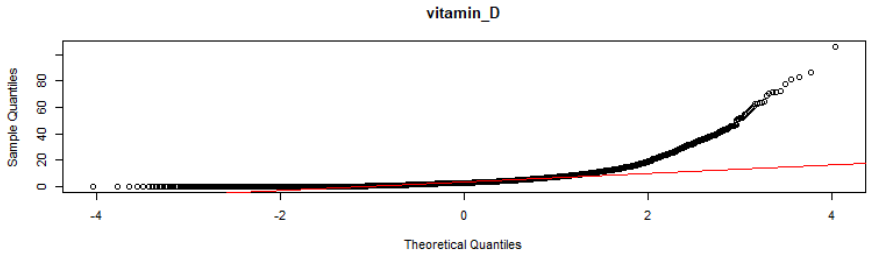


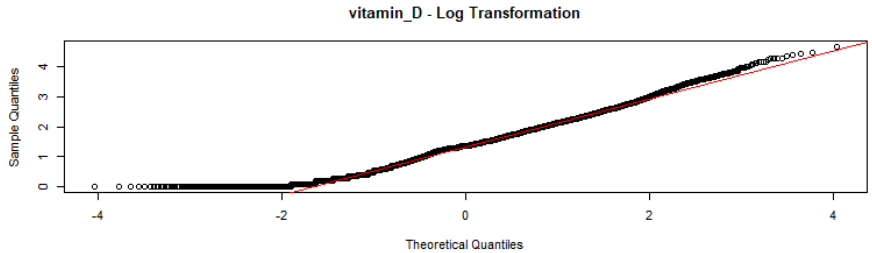


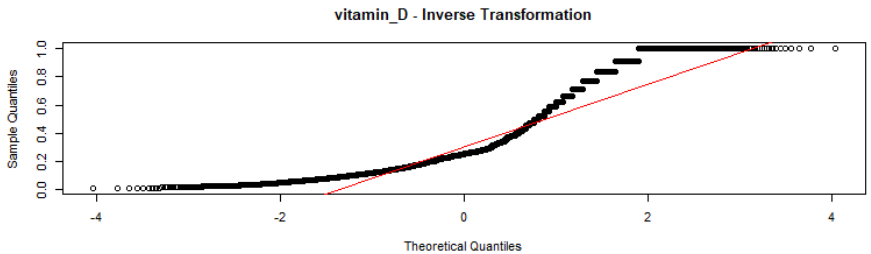


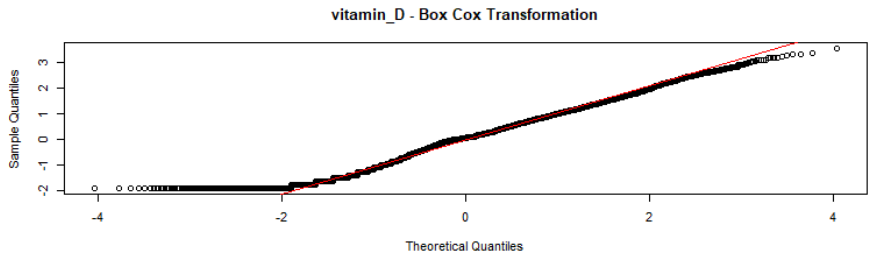
Univariate Normality: Vitamin D

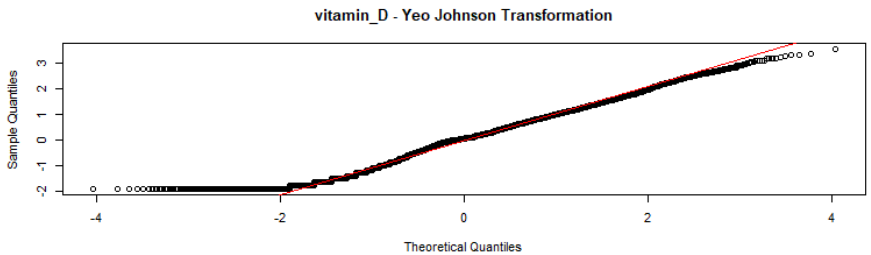
- Assessed normality of Vitamin D using Q-Q plots with:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson transformations





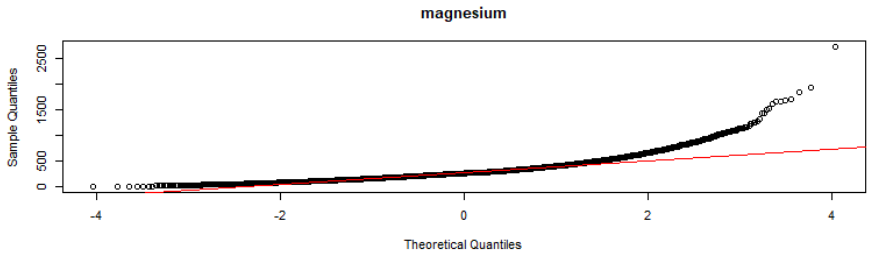


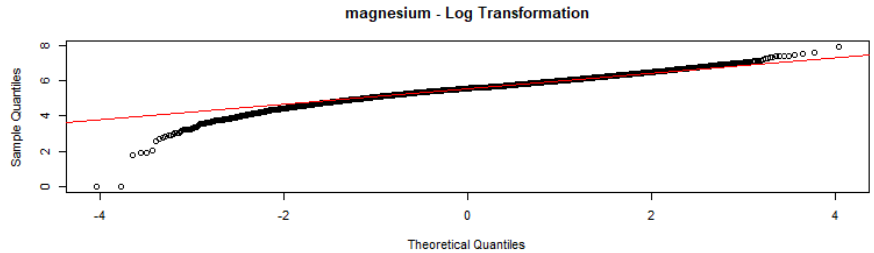


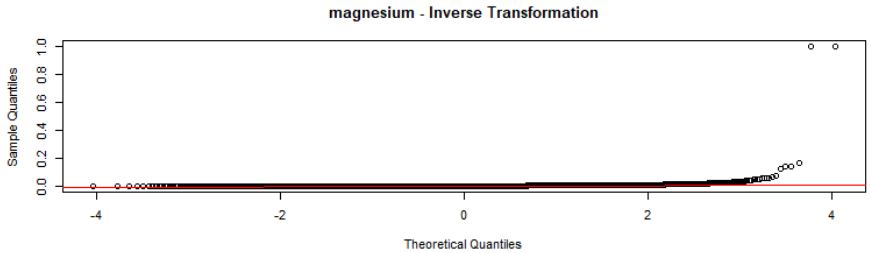


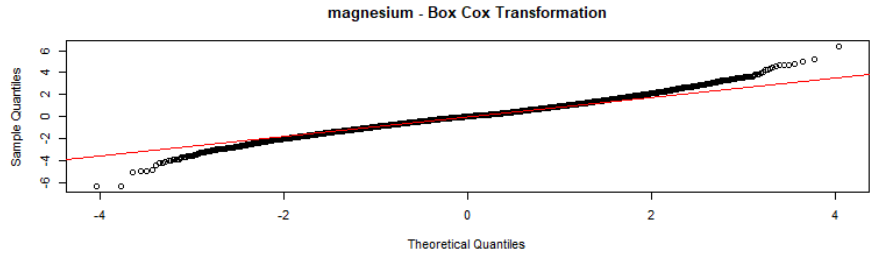
Univariate Normality: Magnesium

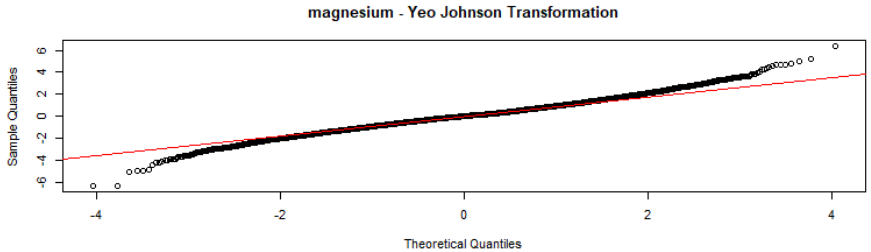
- Q-Q plots used to assess normality with:
  - Raw, Log, Inverse, Box-Cox, and Yeo-Johnson transformations





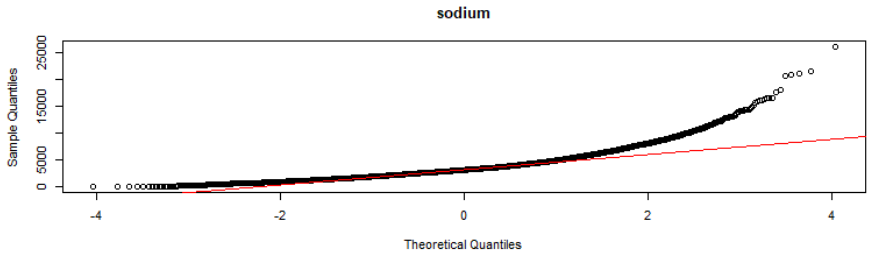


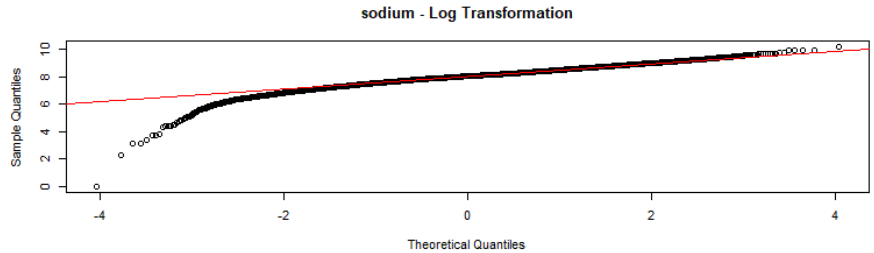


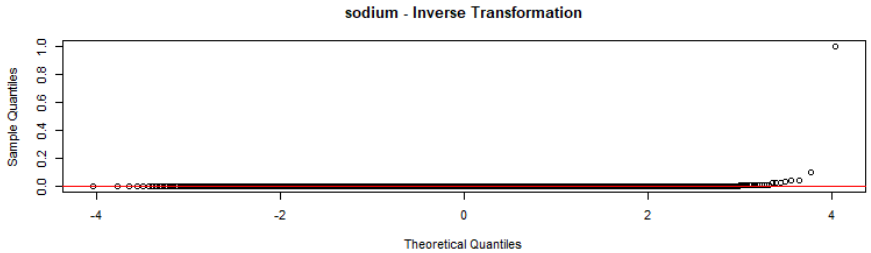


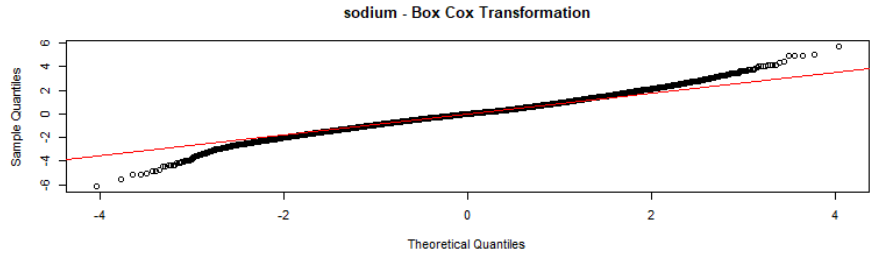
Univariate Normality: Sodium

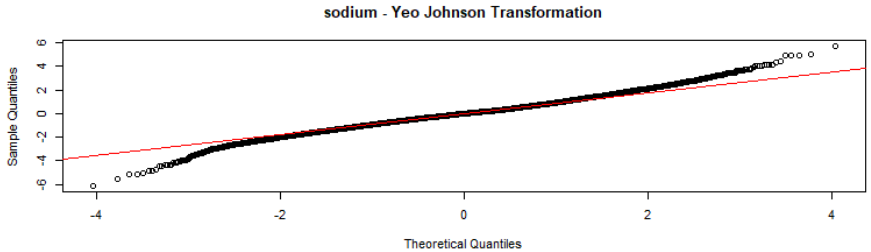
- Q-Q plots analyzed under five transformations:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson





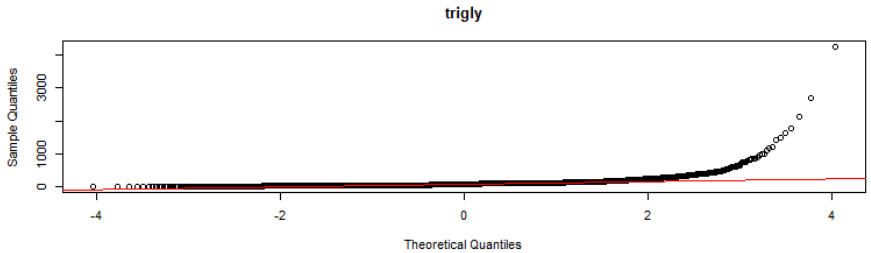


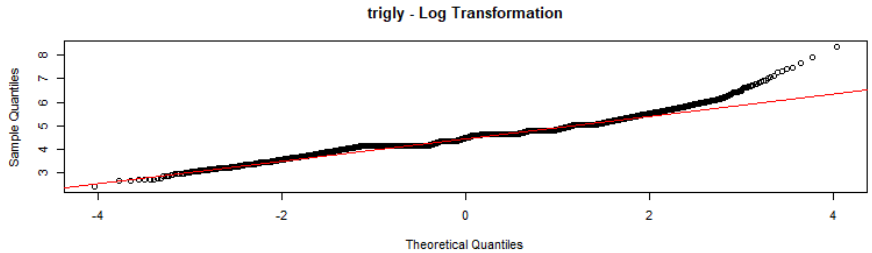


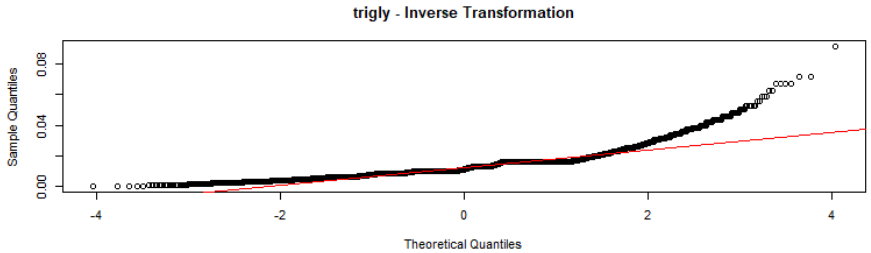


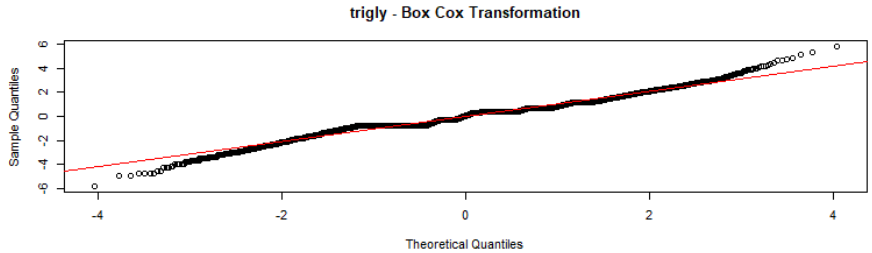
Univariate Normality: Triglycerides

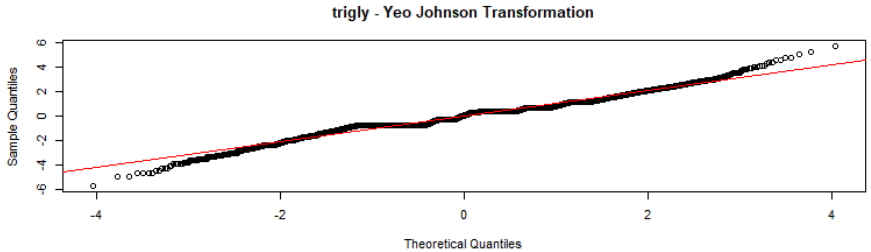
- Q-Q plots were analyzed using:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson transformations





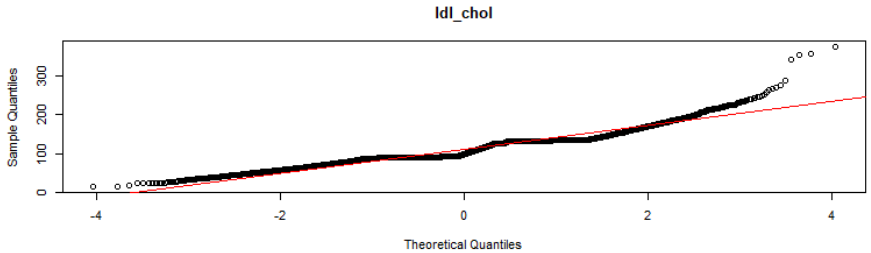


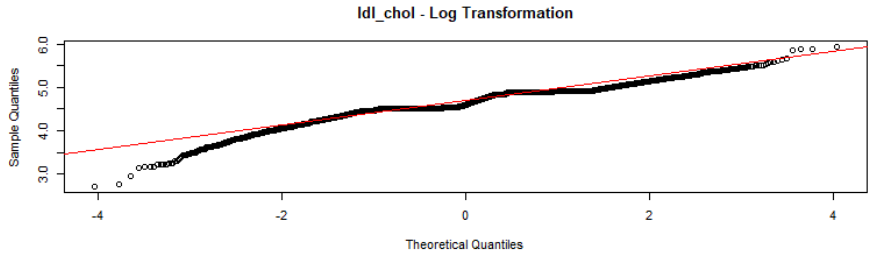


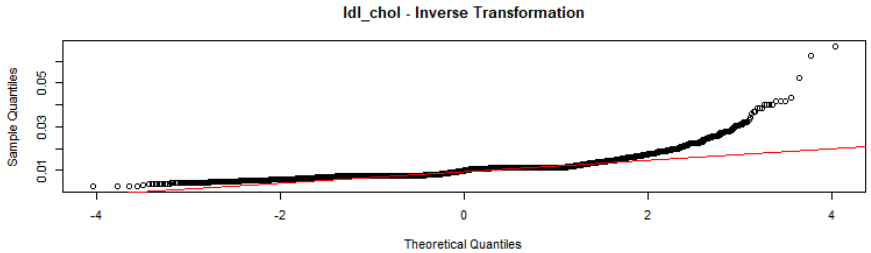


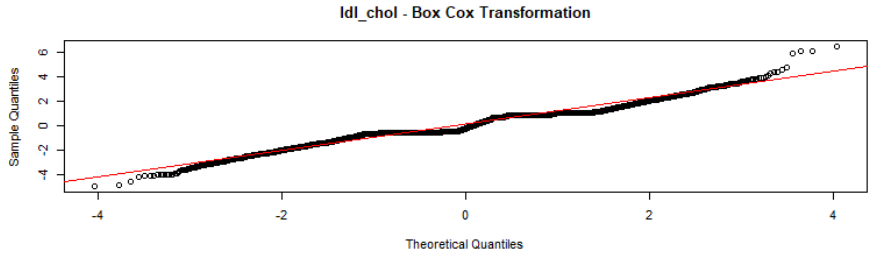
Univariate Normality: LDL Cholesterol

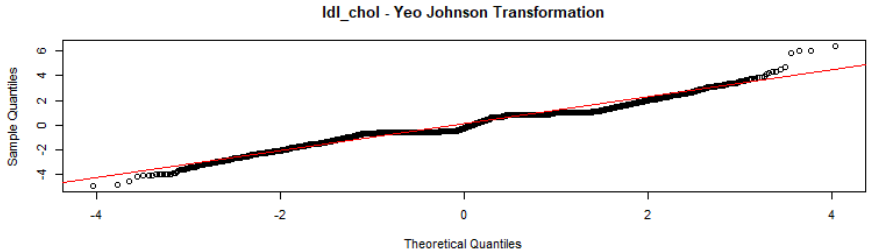
- Assessed normality using Q-Q plots for:
  - Raw, Log, Inverse, Box-Cox, and Yeo-Johnson transformations





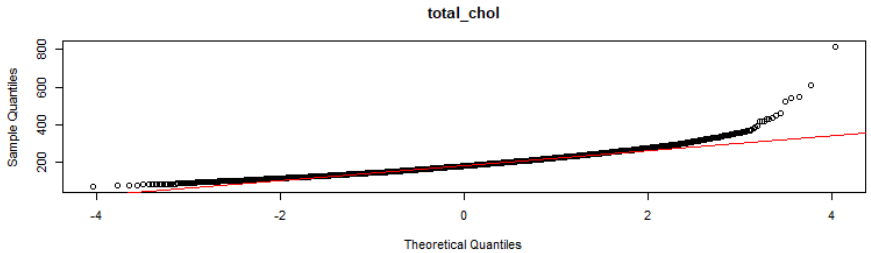


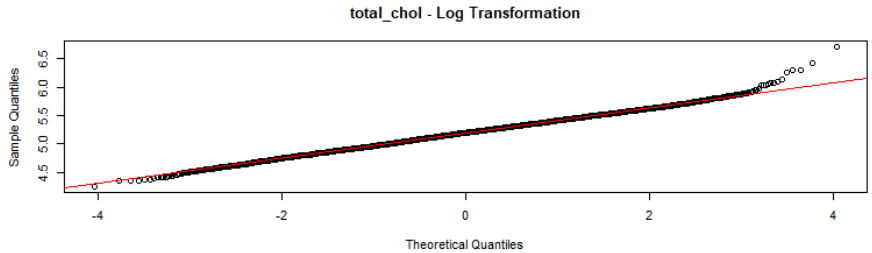


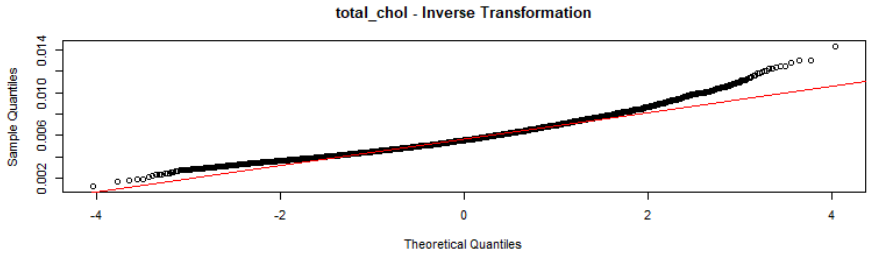


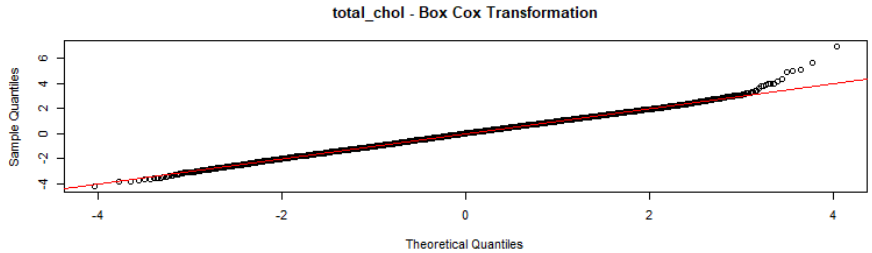
Univariate Normality: Total Cholesterol

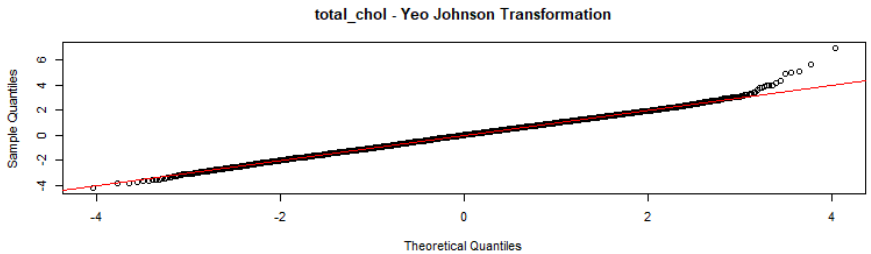
- Q-Q plots assessed using:
  - Raw, Log, Inverse, Box-Cox, Yeo-Johnson transformations





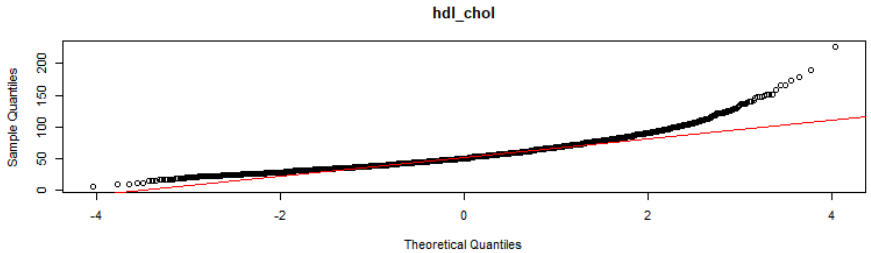


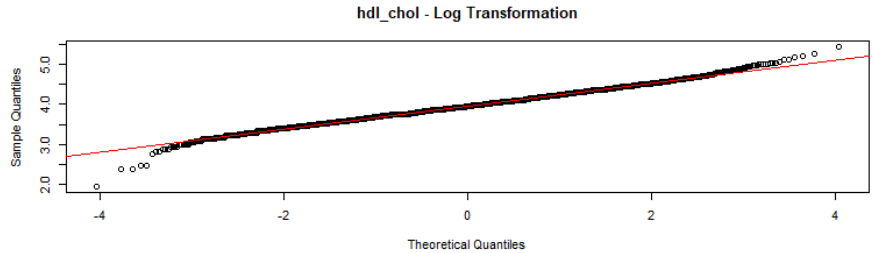


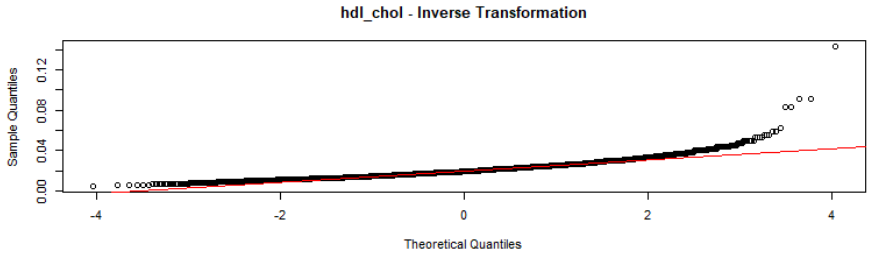


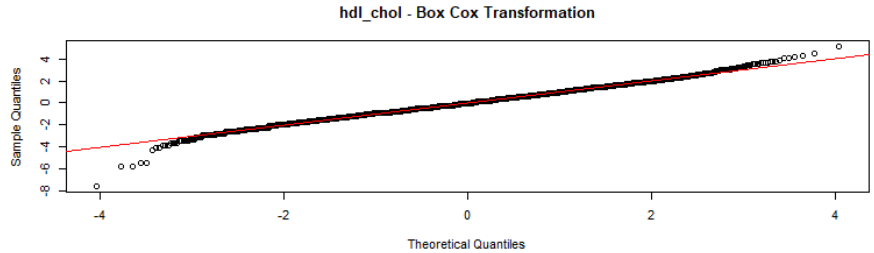
Univariate Normality: HDL Cholesterol

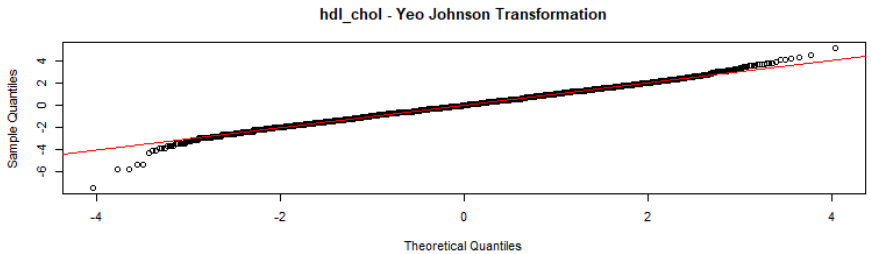
- Q-Q plots analyzed under:
  - Raw, Log, Inverse, Box-Cox, and Yeo-Johnson transformations





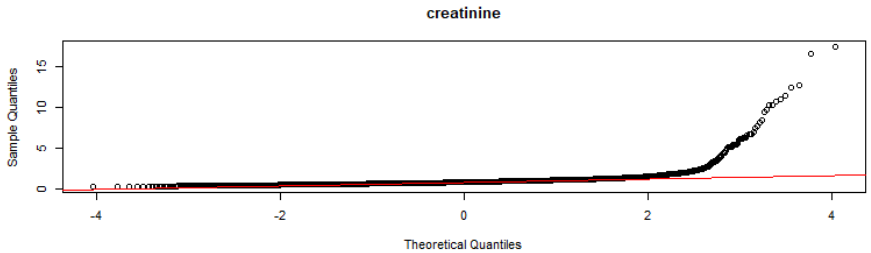


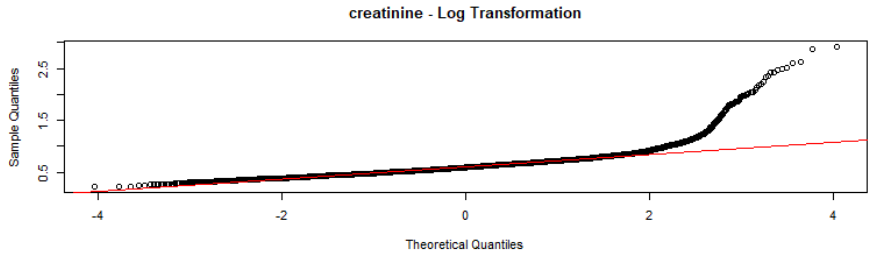


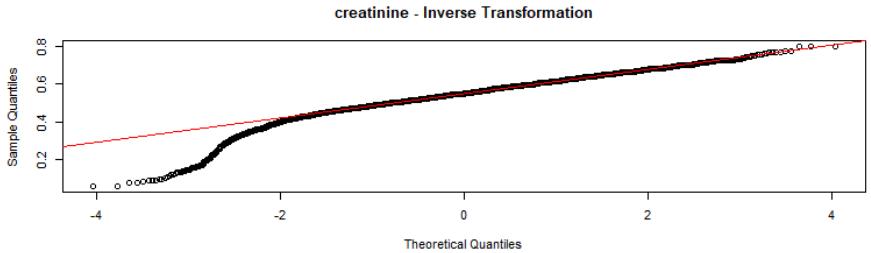


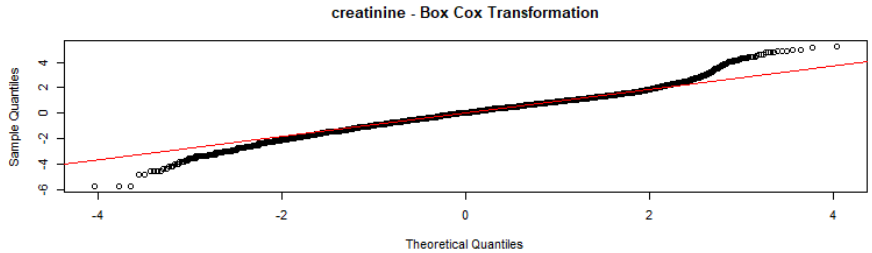
Univariate Normality: Creatinine

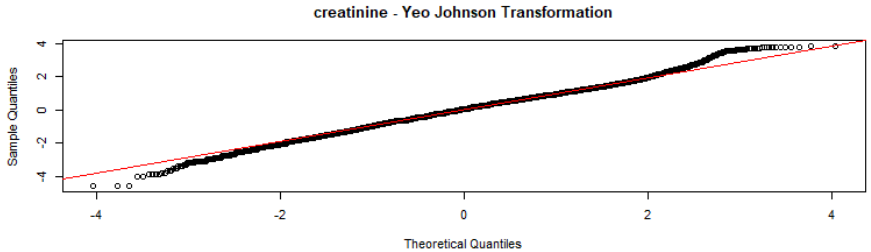
- Q-Q plots analyzed under:
  - Raw, Log, Inverse, Box-Cox, and Yeo-Johnson transformations





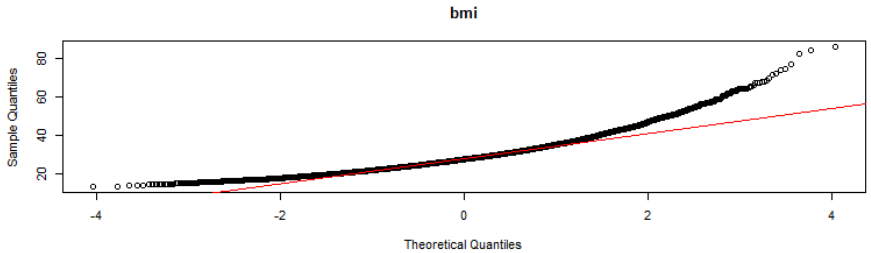


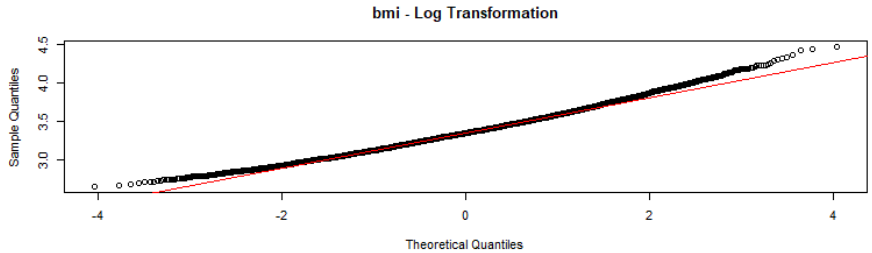


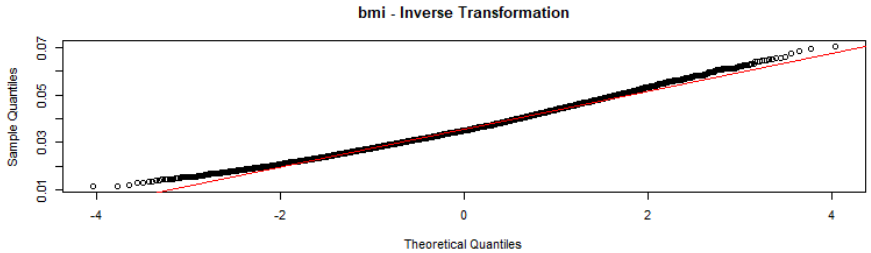


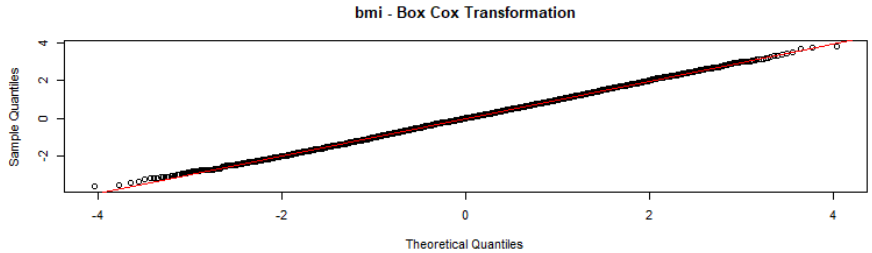
Univariate Normality: BMI

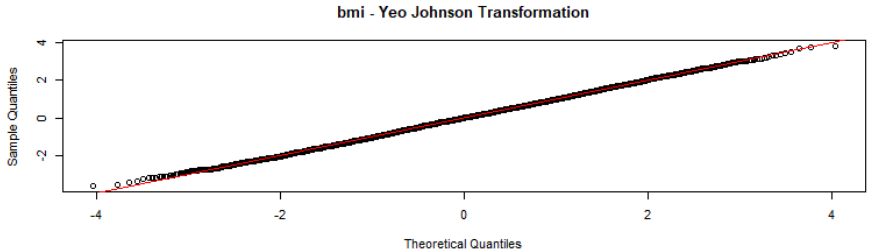
- Q-Q plots analyzed under:
  - Raw, Log, Inverse, Box-Cox, and Yeo-Johnson transformations





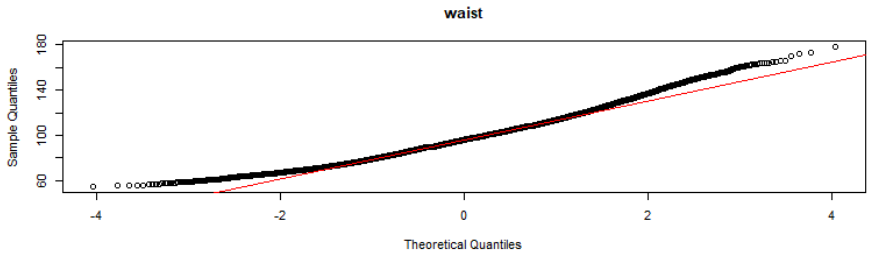


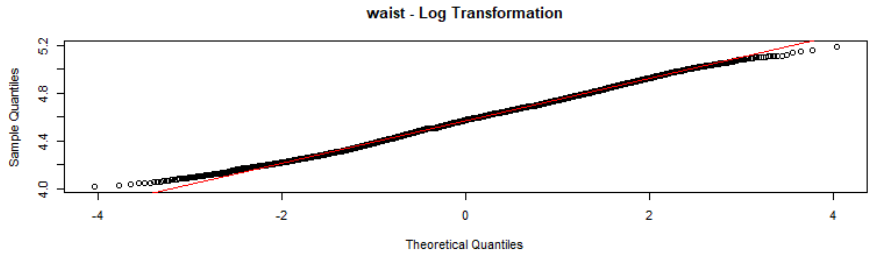


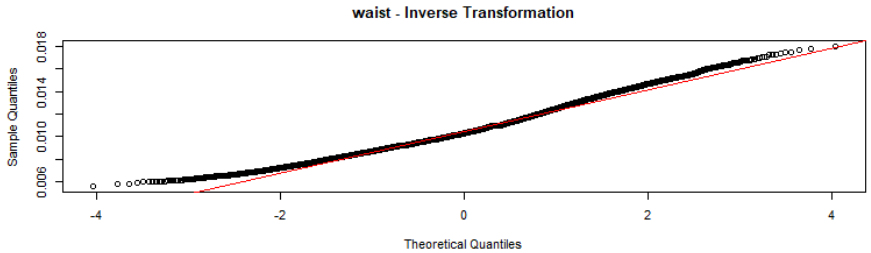


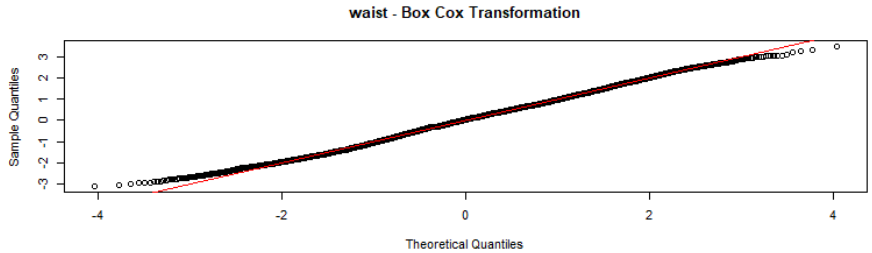
Univariate Normality: Waist Circumference

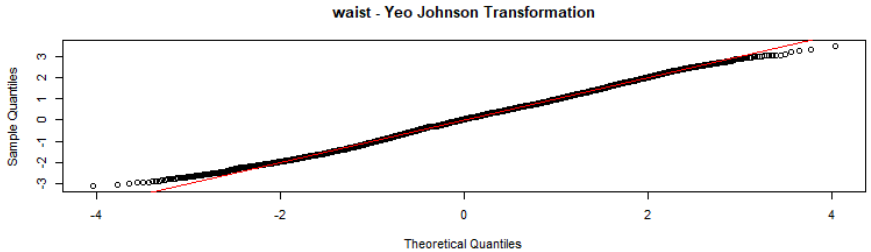
- Q-Q plots analyzed under:
  - Raw, Log, Inverse, Box-Cox, and Yeo-Johnson transformations











Final Comments on Normality Assumptions

- **Univariate Normality:**
  - **BMI, Waist:** After applying the Box-Cox transformation, Q-Q plots align closely with the diagonal, suggesting approximate normality.
  - Since all chosen variables are positive, the **Yeo-Johnson** transformation produces results nearly identical to the Box-Cox transformation.
- Only two variables clearly satisfy the univariate normality assumption.
- **Bivariate Normality:**
  - Analysis using scatterplots and Mahalanobis distances shows that none of the variable pairs exhibit true bivariate normality.
- **Conclusion:**
  - Bivariate normality is not met, even though marginal distributions appear approximately normal.
  - Despite this, discriminant analysis was conducted.

Variable Selection via LASSO Regularization

Variable Selection via LASSO Regularization

Goal: Identify key predictors for an ordinal diabetes outcome using high-dimensional regularization techniques.

- LASSO used to perform variable selection and shrinkage.
- Applied both standard and spline-enhanced LASSO models.
- Ordinal modeling via continuation ratio framework (`glmnetcr` package in R).

Penalized Continuation Ratio Model (`glmnetcr`)

Why `glmnetcr`?

- Suitable for ordinal outcomes (e.g., None, Mild, Moderate, Severe).
- Avoids loss of information from dichotomization.

Model Form:

$$\text{logit}\left(P(Y = k \mid Y \leq k, X = x)\right) = \alpha_k + \beta_k^\top x$$

- Fitted with common slope parameters β across categories.
- Optimal λ chosen via AIC and BIC.

Standard LASSO Implementation

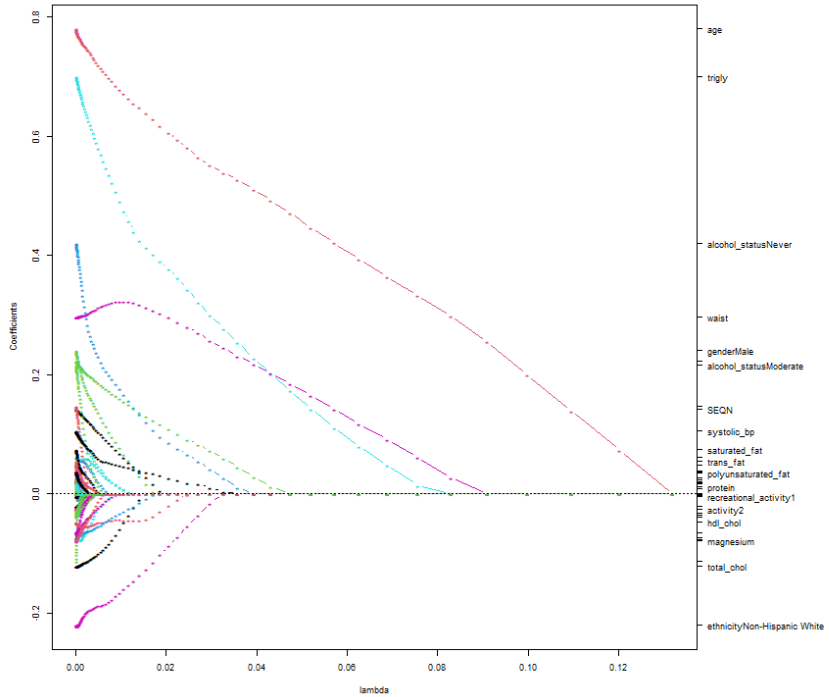


Figure: LASSO coefficient paths for variable selection

Variable Elimination: Categorical Variables

- **Household Size:** All coefficients ≈ 0.
- **Activity Levels:**
  - Activity1: $\beta = 0$
  - Activity2: $\beta = -0.0146$ (negligible)

Variable Elimination: Continuous Variables

Table: Discarded Continuous Variables

Variable	**Coefficient**	**Reason**
Sedentary Minutes	0	Collinear with activity ($r = -0.82$)
Total Cholesterol	-0.116	Derived: $TC = LDL + HDL + \frac{TRIG}{5}$
Creatinine	-0.0081	Low relevance to diabetes
Diastolic BP	-0.0582	Outperformed by systolic BP

Spline-Enhanced LASSO Modeling

- Applied B-spline expansions to capture non-linearity.

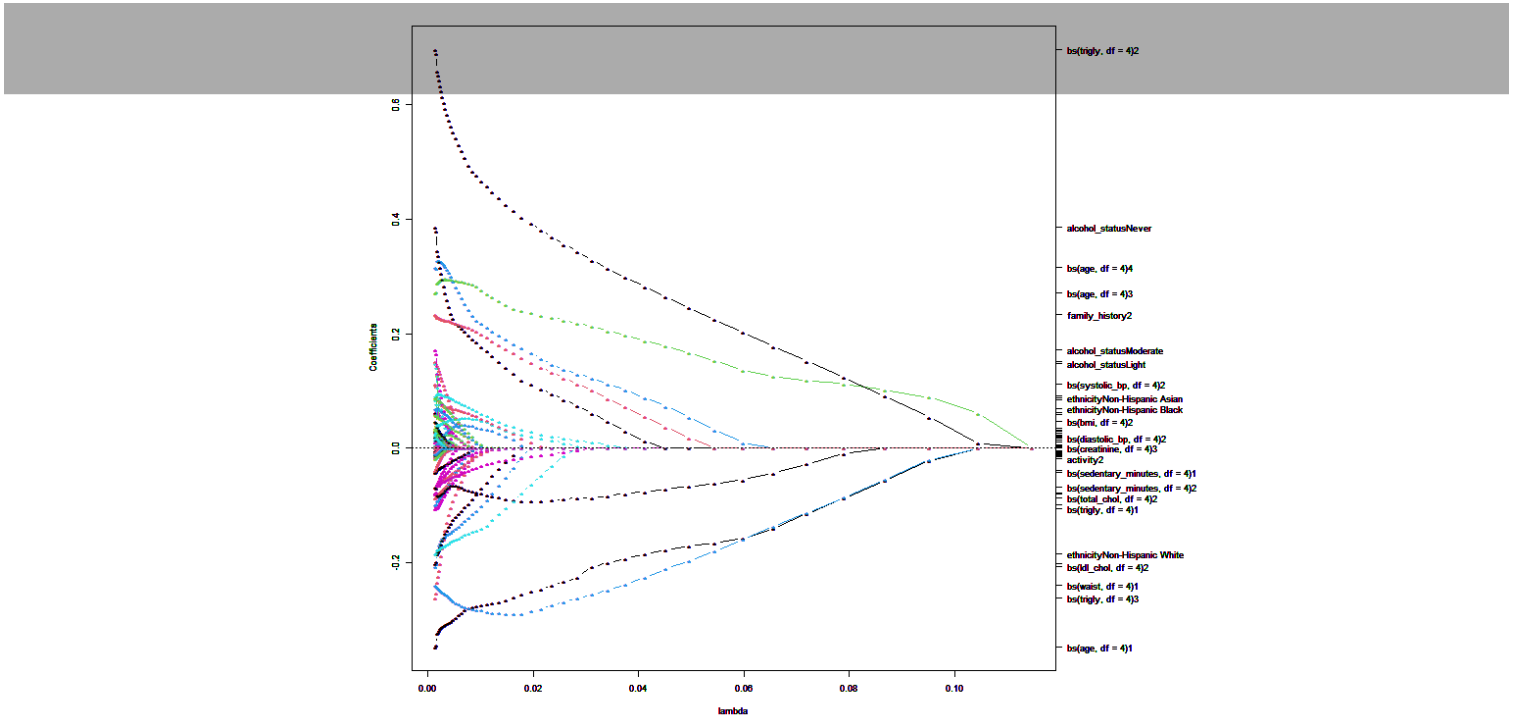


Figure: LASSO paths with spline terms

Key Non-linear Discoveries

Carbohydrates:

- $\beta_{carbs1} = 0.0204$, $\beta_{carbs2} = 0.0169$

Sugar Intake:

- $\beta_{sugar2} = 0.0181$, $\beta_{sugar3} = 0.0190$

Sedentary Minutes:

- $\beta_{sed4} = 0.5319$ (significant)
- Removed due to high collinearity and clinical preference

Modified Variables from LASSO Results

Recreational Activity:

- Originally: 3 levels (0, 1, 2)
- LASSO Results: $\beta_{rec1} = 0$, $\beta_{rec2} = 0.0238$
- Recoded as binary:
  - 0 = None/Moderate
  - 1 = High

Discarded Variables: Summary (1/3)

Categorical / Ordinal Variables

- **Household Size:** All coefficients ≈ 0
- **Activity Levels:** Insignificant (Activity1 = 0, Activity2 = -0.0146)
- **Original Recreational Activity:** Recoded to binary (Rec1 = 0)

Continuous Variables

- **Sedentary Minutes**
  - $\beta_{\text{sed4}} = 0.532$, but excluded due to:
    1. High collinearity with activity ($VIF = 12.4$)
    2. Preference for measures promoting active behavior

Discarded Variables: Summary (2/3)

- **Biochemical Measures**
  - **Total Cholesterol:**
    - Computed as $LDL + HDL + \frac{TRIG}{5}$
    - Redundant, $\beta = -0.116$
  - **Creatinine:** Low relevance for diabetes ($\beta = -0.0081$)
- **Blood Pressure**
  - **Diastolic BP:** Less predictive than systolic (0.0582 vs 0.0879)

Discarded Variables: Summary (3/3)

- **Nutrients and Micronutrients**
  - **Vitamin D:** Insignificant in both linear and spline terms
  - **Magnesium:** Minimal spline effect ($|\beta_{\text{max}}| = 0.0524$)
  - **Sodium:** No clear pattern across categories
  - **Protein:**
    - Boxplot shows minimal variation
    - Inconsistent spline results ($\beta_{\text{prot3}} = -0.235$)

Homogeneity Check of Covariance Matrices

Homogeneity Check of Dispersion Matrices

Objective: Assess whether the Dispersion matrices differ across diabetes status groups.

Methods Used:

- Nonparametric Permutation Test (based on Frobenius norm)
- Box's M Test (parametric, assumes multivariate normality)

Why it Matters:

- Key assumption for Linear Discriminant Analysis
- Covariance structure affects classification and interpretability

Permutation Test: Methodology

Step 1: Group Covariance Matrices

$$\hat{\Sigma}_k = \frac{1}{n_k - 1}(\mathbf{X}_k - \bar{\mathbf{X}}_k)^\top (\mathbf{X}_k - \bar{\mathbf{X}}_k)$$

Step 2: Test Statistic

$$T = \sum_{1 \le i < j \le K} \|\hat{\Sigma}_i - \hat{\Sigma}_j\|_F^2$$

Frobenius Norm:

$$\|A\|_F = \sqrt{\text{tr}(A^\top A)}$$

Permutation Test: Results

Procedure (P = 10,000 replicates):

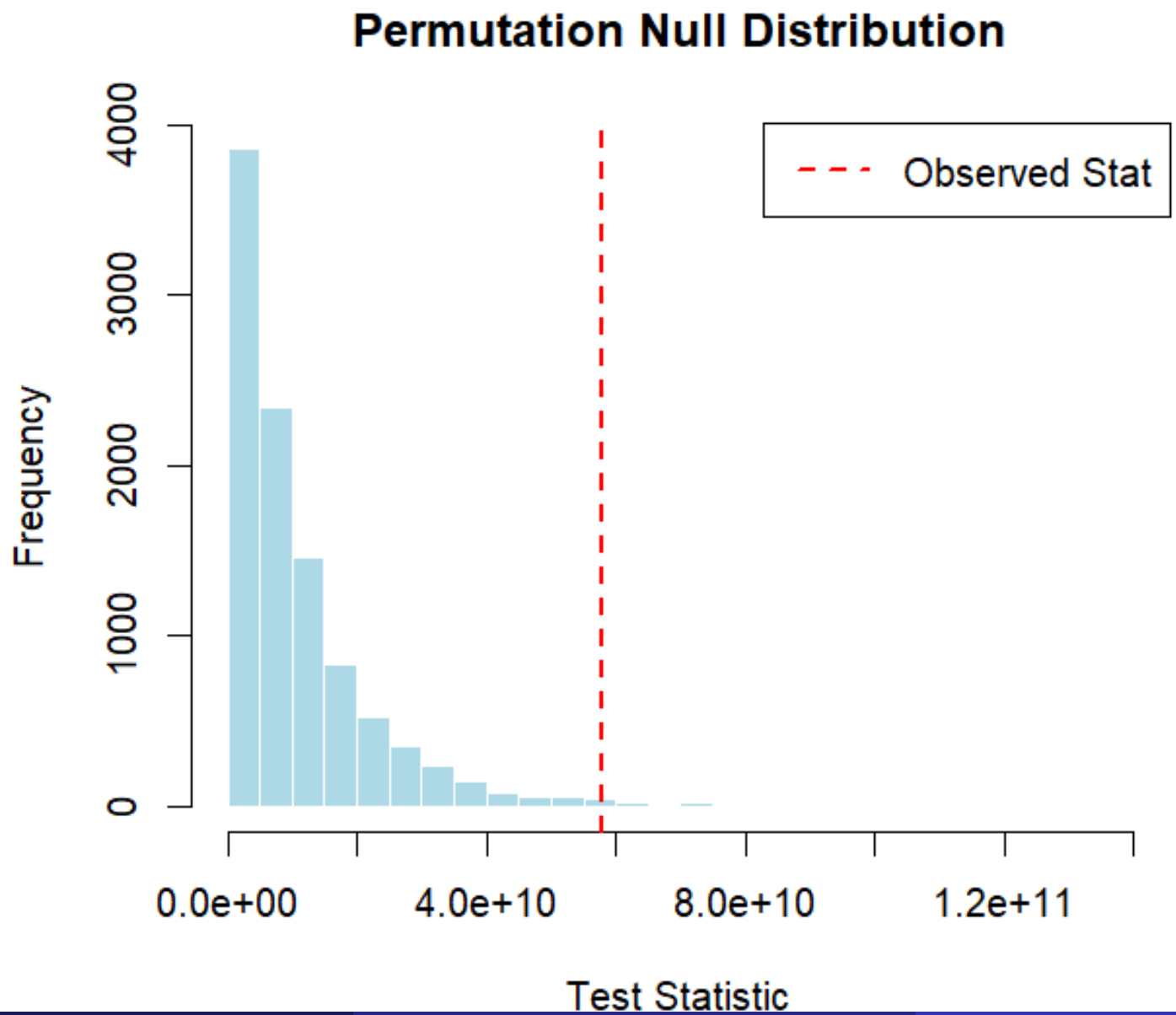
1. Compute T_{obs} using original labels
2. Shuffle labels, recompute $T^{(p)}$ for each permutation
3. Compute empirical p-value:

$$p = \frac{\sum I(T^{(p)} \geq T_{\text{obs}})}{P}$$

Results:

- Observed $T_{\text{obs}} = 5.79 \times 10^{10}$
- p-value $= 0.0076$
- $\Rightarrow$ Strong evidence against homogeneity

Empirical Distribution of Test Statistic



Box's M Test

For K groups with p variables, let n_k be the sample size and S_k the sample covariance matrix for group k. The pooled covariance matrix is:

$$S_p = \frac{\sum_{k=1}^{K}(n_k - 1)S_k}{N - K} \tag{1}$$

where $N = \sum_{k=1}^{K} n_k$. The test statistic is:

$$M = \gamma \left[(N - K) \ln |S_p| - \sum_{k=1}^{K} (n_k - 1) \ln |S_k| \right] \tag{2}$$

Box's M Test: Classical Approach

$$\gamma = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(K-1)} \left(\sum_{k=1}^{K} \frac{1}{n_k - 1} - \frac{1}{N-K} \right) \tag{3}$$

Under H_0, $M \sim \chi^2_v$ where $v = \frac{1}{2}p(p+1)(K-1)$.

Results:

- $M = 13,579$, df $= 210$
- p-value $< 2.2 \times 10^{-16}$
- Confirms permutation test findings with overwhelming evidence

Interpretation and Diagnostic Insights

Both Tests Conclude:

- Covariance matrices differ significantly across groups
- LDA may be inappropriate due to violated assumptions
- Suggests use of QDA or models with group-specific dispersion

Caveats:

- Box's M is highly sensitive in large samples or large number of variables
- Permutation test more robust, but computationally intensive

Discriminant Analysis

Discriminant Analysis

Objective: Predict diabetes status using multivariate discriminant functions.

Challenges:

- Class imbalance (Normal: 53%, Prediabetic: 35%, Diabetic: 12%)
- Heterogeneous covariance structures (QDA appropriate)

Strategy:

- Use Quadratic Discriminant Analysis (QDA) & Linear Discriminant Analysis (LDA)
- Apply inverse-frequency priors to handle imbalance

Addressing Class Imbalance

Original Distribution:

- Normal: 53%, Prediabetic: 35%, Diabetic: 12%

Inverse-Frequency Priors:

$$\text{Priors} = \left(\frac{1}{n_0}, \frac{1}{n_1}, \frac{1}{n_2}\right) / \sum \left(\frac{1}{n_i}\right) = (0.15, 0.23, 0.62)$$

Benefit:

- Boosts contribution of minority classes, especially diabetics
- Mitigates majority class bias

Quadratic Discriminant Analysis (QDA)

Discriminant Function:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

Key Assumptions:

- Each class has its own covariance matrix Σ_k
- Allows for nonlinear (quadratic) decision boundaries

Why QDA?

- Suitable when homogeneity of covariances is violated (as established)

QDA Results: Confusion Matrix

Actual / Predicted	Normal	Prediabetic	Diabetic
Normal	1925	492	39
Prediabetic	598	539	80
Diabetic	498	938	596

High Recall for Diabetics: 83.4%
Low Sensitivity for Prediabetics: 27.4%

QDA Multiclass ROC Curve

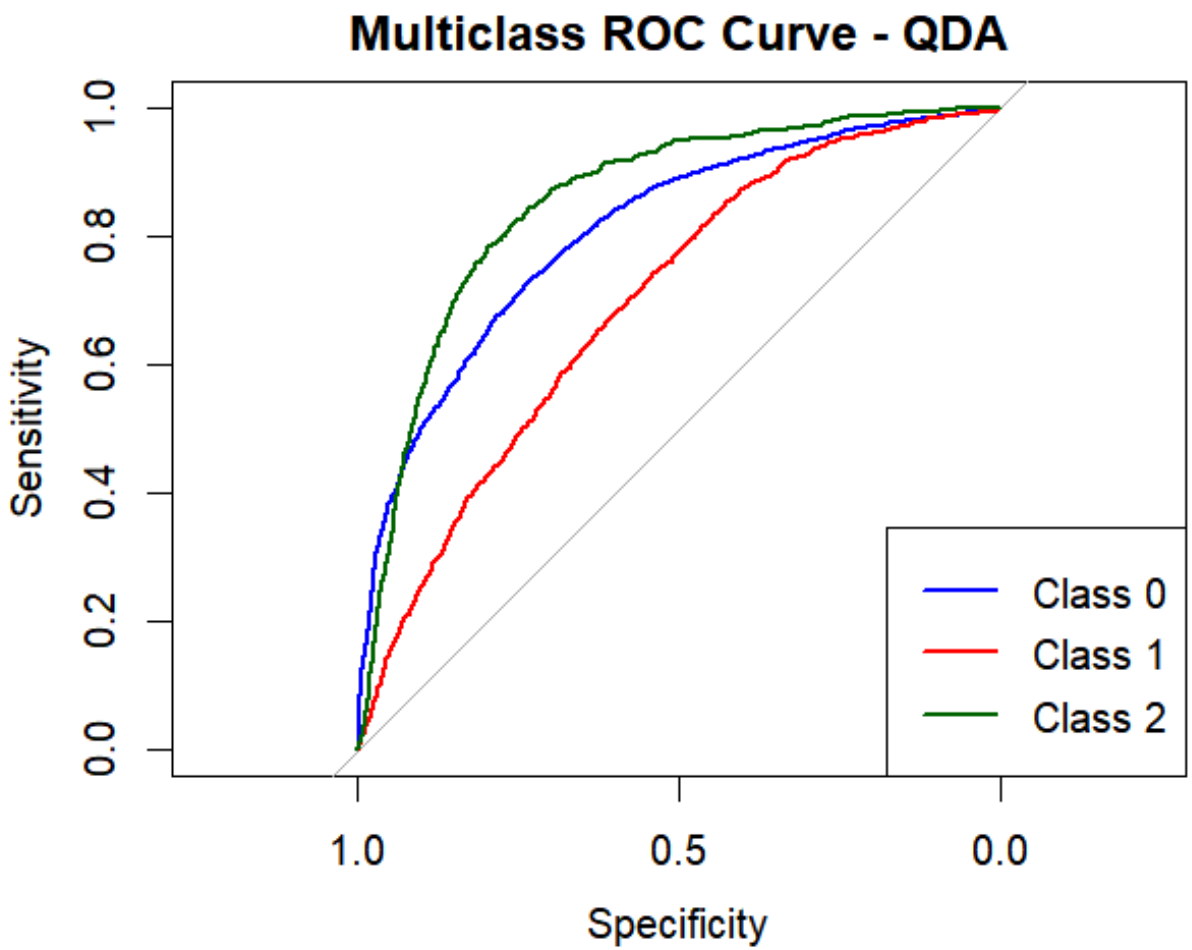


Figure: ROC Curves by Class (QDA)

QDA Multiclass ROC Curve: Observations

Observations:

- Class 2 (Diabetic) — steep, near-perfect curve
- Class 0 (Normal) — strong performance
- Class 1 (Prediabetic) — lower separability

QDA Precision-Recall Curve

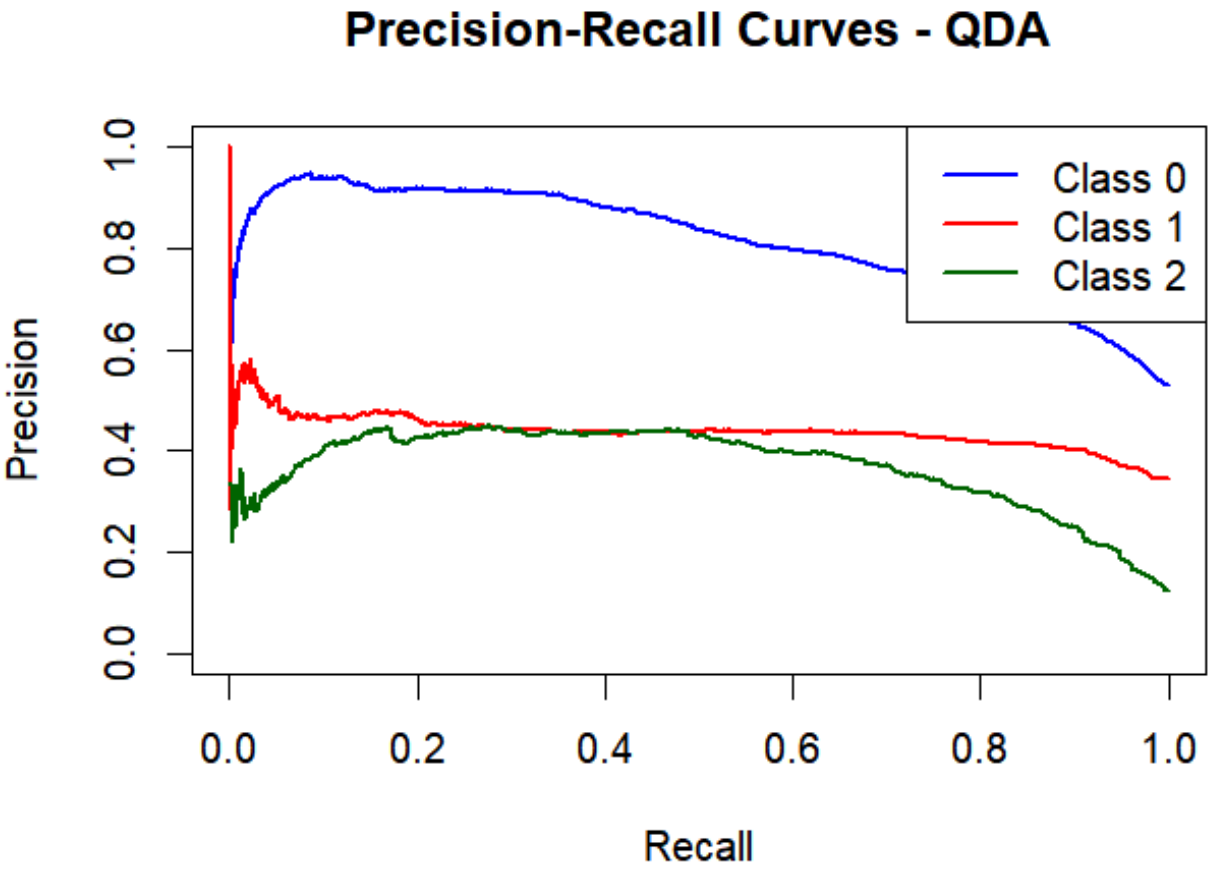


Figure: Precision-Recall Curves by Class (QDA)

QDA Precision-Recall Curve: Observations

Observations:

- Excellent Class 2 precision/recall
- Class 1 precision drops at high recall — overlapping boundaries
- Overall strong positive predictive power for diabetic detection

QDA Precision-Recall Curve

Key Takeaways:

- Precision and recall balanced across classes
- Class 2 (Diabetic) maintains strong precision despite lower recall
- Precision drop-offs in Classes 0 and 1 reflect misclassification tradeoffs

Model Insights

Strengths:

- Diabetics identified with high recall and AUC = 0.91
- Strong specificity for normal class (83.4%)

Limitations:

- Poor performance on prediabetic class (PPV = 44.3%)
- 938 prediabetics misclassified as diabetic — risk of overtreatment

Clinical Implication:

- QDA effective for identifying severe cases (diabetic)
- Less reliable in capturing nuanced prediabetic profile

Linear Discriminant Analysis (LDA)

- LDA assumes a shared covariance matrix Σ across all classes.
- Leads to linear decision boundaries between classes.

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Confusion Matrix

Actual / Predicted	Normal	Prediabetic	Diabetic
Normal	2373	767	73
Prediabetic	613	1067	423
Diabetic	35	135	219

LDA – Performance Insights

Advantages:

- Higher overall accuracy (64.1%) compared to QDA (53.6%)
- Significantly improved sensitivity for the prediabetic class (54.2% vs QDA's 27.4%)

Limitations:

- Diabetic class is severely under-identified (Recall = 30.6%)
- Substantial false positives in prediabetic predictions (767 normal individuals misclassified)

LDA – ROC Curve

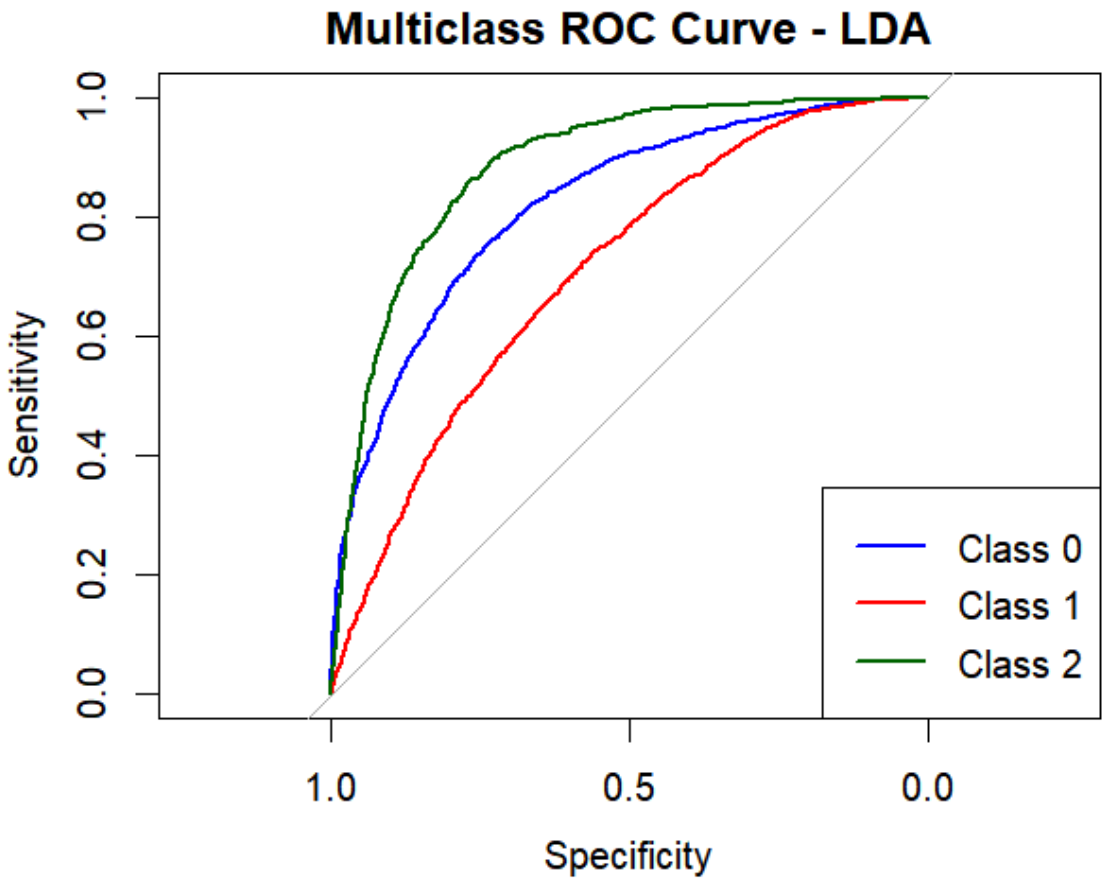


Figure: Multiclass ROC Curve for LDA

LDA – ROC Curve: Observations

Observations:

- Class 2 shows better separation.
- Class 1 exhibits lowest AUC.
- Overall, moderate sensitivity and specificity.

LDA – Precision-Recall Curve

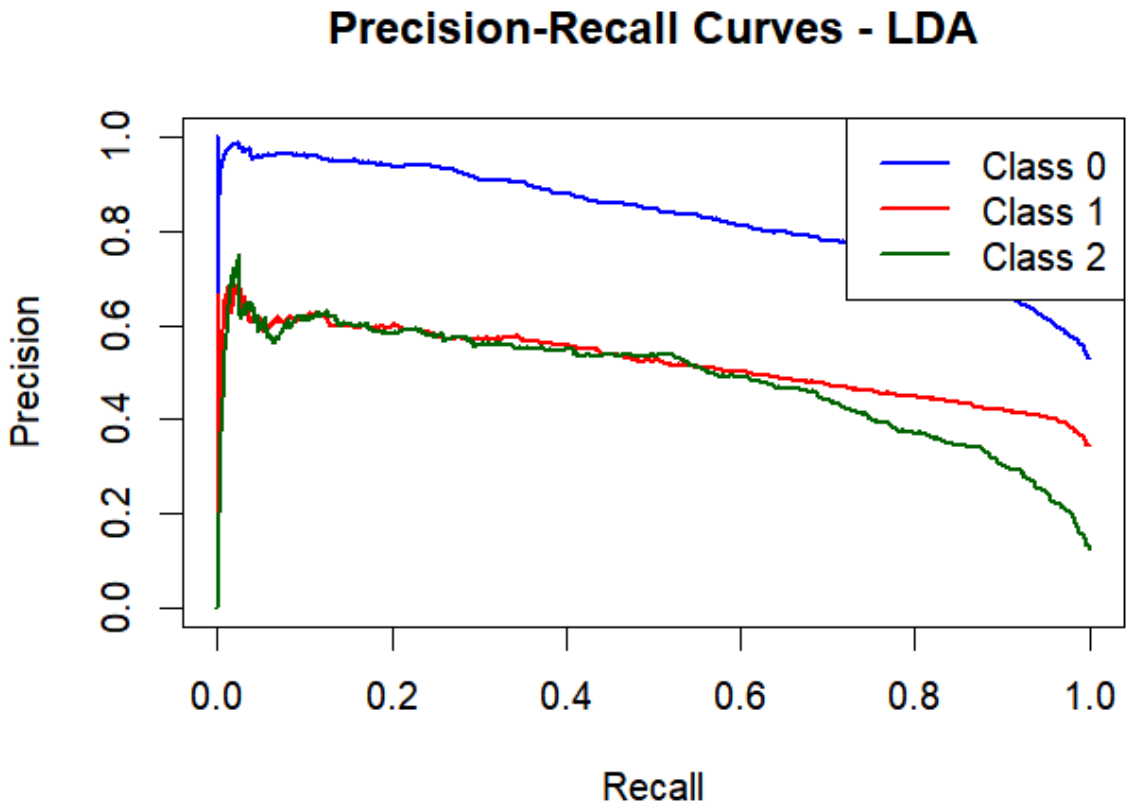


Figure: Precision-Recall Curve for LDA

LDA – Precision-Recall Curve: Observations

Observations:

- Class 0 has strongest precision-recall balance.
- Drop-offs in precision for Classes 1 and 2 suggest moderate confusion.
- Adequate performance with room for improvement in ambiguous cases.

LDA – Mathematical Interpretation

- LDA assumes shared covariance across classes.
- This assumption leads to linear decision boundaries.
- Fails to capture nonlinear patterns, e.g., complex lipid-BMI interactions in diabetic patients.

10-Fold Cross Validation

Goal: Evaluate robustness model using stratified 10-fold cross-validation.

Procedure:

- Used ‘caret::createFolds()‘ to generate 10 folds, preserving class distribution.
- Iteratively trained LDA and QDA models on 9 folds and tested on the remaining one.
- Repeated the process for all 10 folds.

10-Fold Cross Validation

- **Stratification**: Maintained original class proportions in each fold:
  - Normal: 53%, Prediabetic: 35%, Diabetic: 12% ($\pm 1\%$)
- **Comprehensive Evaluation**:
  - Stored class predictions and posterior probabilities
  - Aggregated results across folds
  - Generated ROC and Precision-Recall curves from posterior data
- More reliable than a single train-test split
- Reduces variance in performance metrics

QDA Cross-Validation

Confusion Matrix Summary

Table: Cross-Validated Confusion Matrix for QDA (Avg. Across 10 Folds)

Actual / Predicted	Normal	Prediabetic	Diabetic
Normal	7758	2581	333
Prediabetic	1951	3017	1035
Diabetic	362	966	1018

Key Observation: Diabetic class recall drops significantly in CV compared to the test set, despite strong performance.

QDA Cross-Validation

Detailed Class Insights

Normal Class

- High precision and stable detection across folds.
- Low diabetic misclassification rate — good separation of healthy cases.

Prediabetic Class

- Most challenging class to identify.
- Significant overlap with both normal and diabetic profiles.
- Risk of overtreatment due to false diabetic classification.

Diabetic Class

- Sensitivity reduced from 83.4% to 42.7% in CV.
- Still superior to LDA for high-risk cases.
- Suggests QDA's ability to model non-linear boundaries effectively.

QDA Cross-Validation

Multiclass ROC Curve

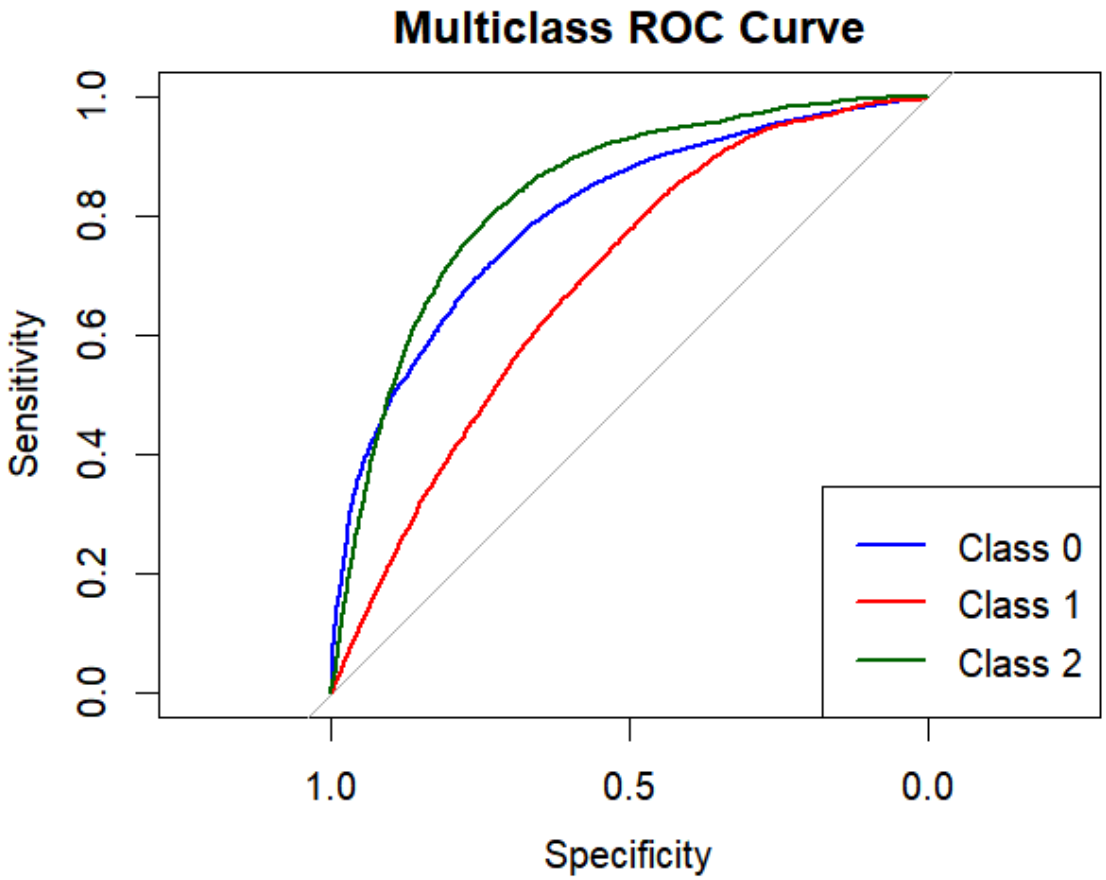


Figure: ROC Curve for QDA (Cross-Validated)

QDA Cross-Validation: Observations

Multiclass ROC Curve

Observations:

- Class 2 (Diabetic) leads in AUC performance.
- Fluctuations suggest moderate sensitivity to fold composition.
- ROC confirms overall robustness of QDA.

QDA Cross-Validation

Precision-Recall Curve

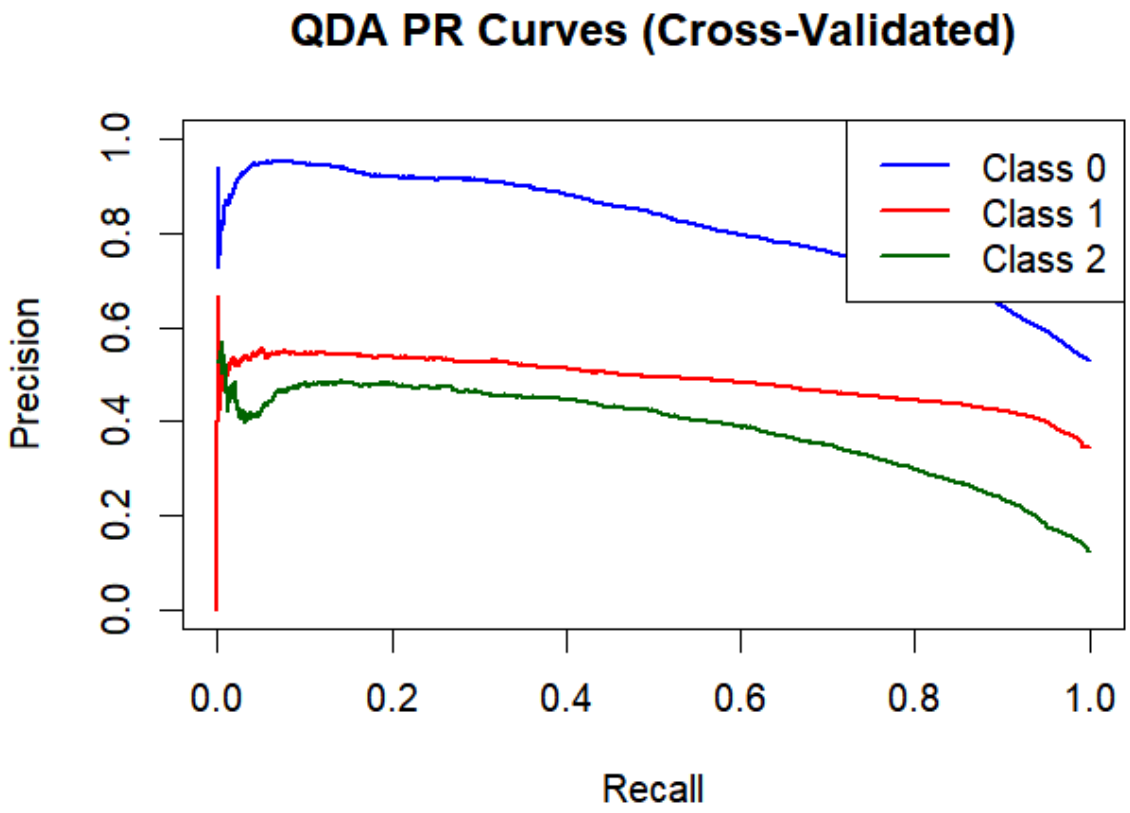


Figure: Precision-Recall Curve for QDA (Cross-Validated)

QDA Cross-Validation: Observations

Precision-Recall Curve

Observations:

- Strong PR performance for Diabetic class.
- Class 1 remains problematic — high confusion remains.
- Curves smoothened by CV, indicating improved generalization.

QDA Cross-Validation

Performance Insights

- **Diabetic AUC**: Remains high (0.91), but sensitivity drops by 33%.
- **Prediabetic Class**: Precision-recall curve drops steeply after recall = 0.6.
- **QDA Generalization**: Strong ROC/PR performance retained, but borderline cases remain difficult.

LDA Cross-Validation

Confusion Matrix Summary

Table: Cross-Validated Confusion Matrix for LDA (Avg. Across 10 Folds)

Actual / Predicted	**Normal**	**Prediabetic**	**Diabetic**
Normal	7976	2577	251
Prediabetic	1971	3525	1381
Diabetic	124	462	754

Observation: Strong normal class classification; underperformance for diabetic class.

Normal Class

- High recall and minimal misclassification into diabetic class.
- Demonstrates strong separation for low-risk profiles.

Prediabetic Class

- Better recall than QDA (53.7% vs 46.0%).
- Still prone to misclassifications into diabetic group.

Diabetic Class

- Diabetic recall reduced to 31.6%.
- Many diabetics misclassified as prediabetics — potential clinical implications.

LDA Cross-Validation

ROC Curve Analysis

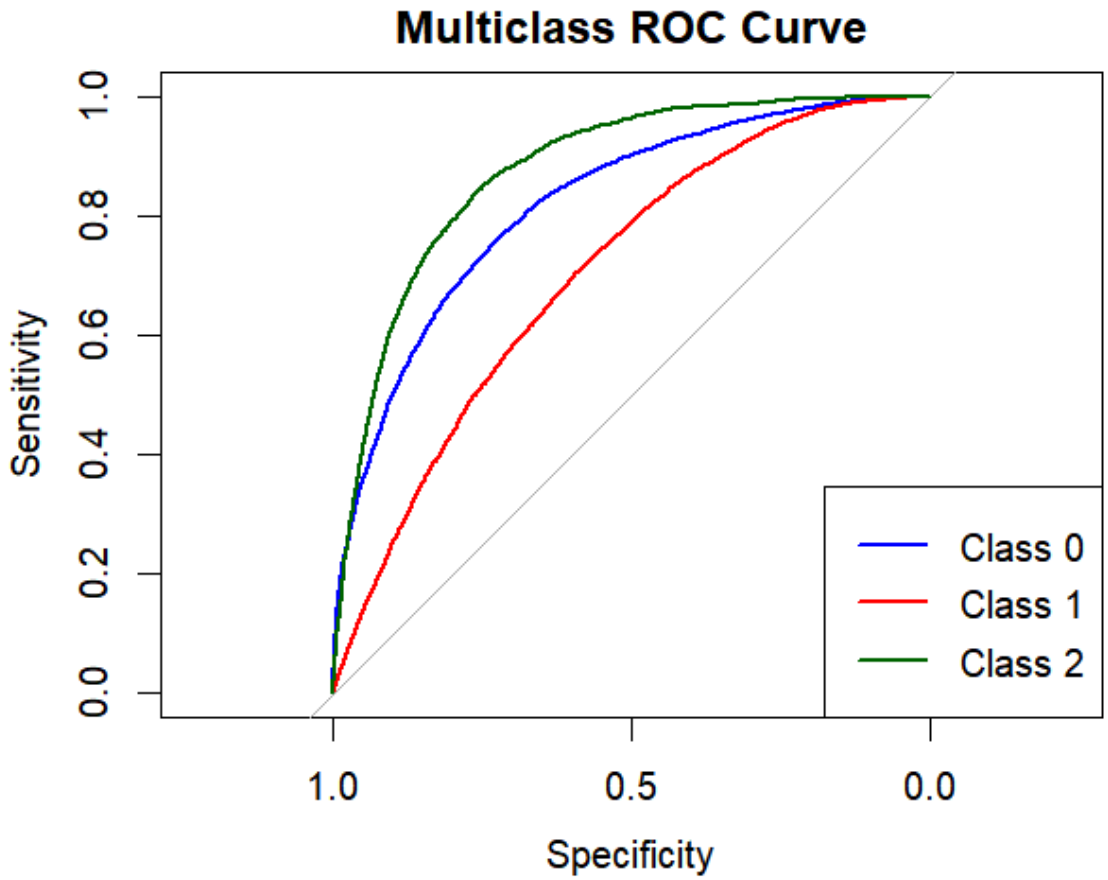


Figure: Multiclass ROC Curve for LDA (Cross-Validated)

LDA Cross-Validation: Observations

ROC Curve Analysis

Observations:

- ROC curves show smoother, more generalizable class separation.
- Slight improvements for Class 2 compared to non-CV performance.
- Suggests better generalization and reduced overfitting.

LDA Cross-Validation

Precision-Recall Curve

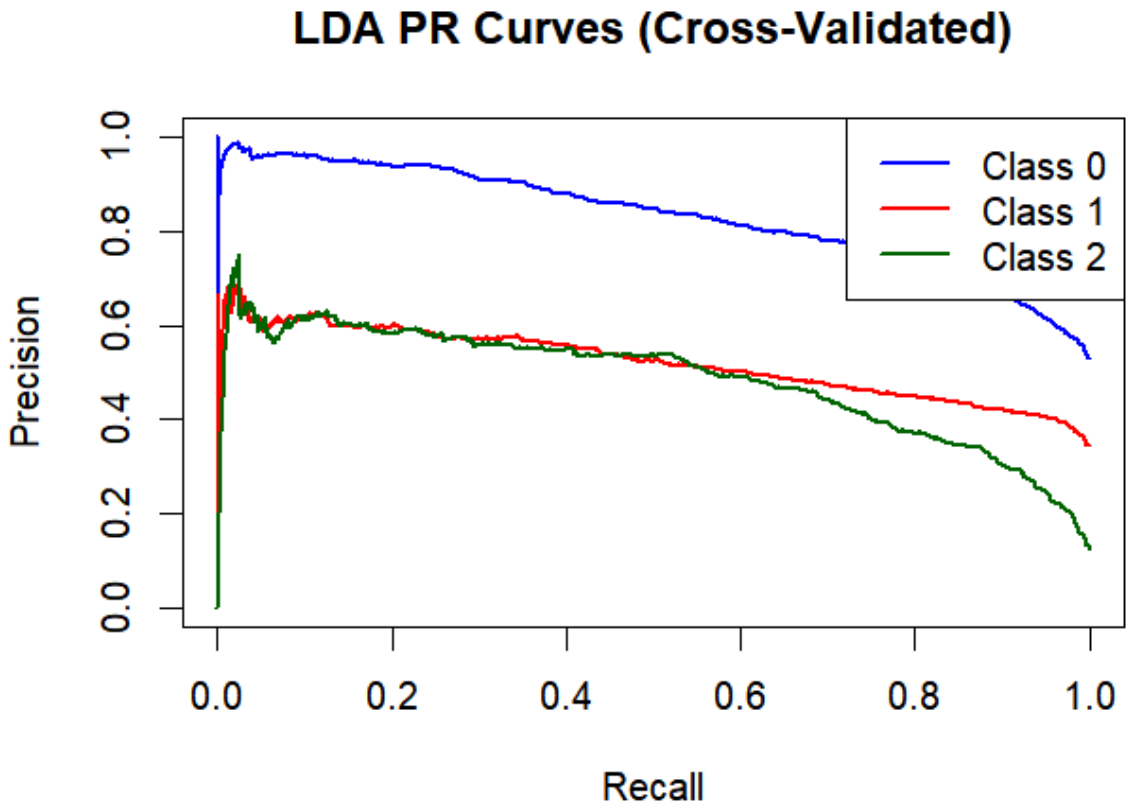


Figure: Precision-Recall Curve for LDA (Cross-Validated)

LDA Cross-Validation: Observations

Precision-Recall Curve

Observations:

- Clear improvements in Class 0 (Normal) precision and recall.
- Class 1 still trails behind but sees smoother curve behavior.
- Cross-validation stabilizes PR performance across classes.

LDA Cross-Validation

Key Takeaways

- **Prediabetic Recall:** 53.7% — stronger than QDA.
- **Diabetic Detection:** Weaker with 31.6% recall.
- **Model Insight:** LDA's linearity suits low-to-moderate risk; struggles at higher complexity boundaries.
- **Cross-Validation Benefits:** Improves calibration and generalization; stabilizes performance curves.

Model Comparison

QDA vs LDA – Performance Metrics

Table: Comparative Performance of QDA and LDA Classifiers

Category	**Metric**	**QDA**	**LDA**
Overall Accuracy	Test Set Accuracy	53.6%	64.1%
	CV Accuracy	62.0%	**64.4%**
Class-Level Performance	Diabetic Recall	**83.4%**	30.6%
	Prediabetic Recall	27.4%	**54.2%**
	Normal Specificity	**80.2%**	68.7%

QDA Advantages

- Best at identifying diabetic cases — crucial for screening.
- Handles complex, non-linear patterns (e.g., lipid ratios).
- Strong specificity in normal population — fewer false positives.

LDA Strengths

- More balanced accuracy — better general performance.
- Doubles recall for prediabetics — better risk stratification.
- Simpler and faster to train; more stable across CV folds.

Clinical Trade-offs

- **QDA:** Lower diabetic miss rate but more prediabetic overdiagnosis.
- **LDA:** Lower diabetic recall but better at managing moderate-risk groups.

Final Assessment

- QDA is preferred when missing a diabetic case is costly.
- LDA is preferred when overall risk calibration and balanced identification are needed.
- Model choice should be guided by clinical priorities — e.g., screening vs. stratification.

Fisher's Discriminant Analysis

Goal: Maximize class separability via projection.

Key Matrices:

- S_B: Between-class scatter — measures class separation
- S_W: Within-class scatter — measures class compactness

Eigenvalue Problem:

$$S_W^{-1} S_B w = \lambda w$$

Result: Projection vectors w maximize inter-class variance.

Fisher's Discriminant Analysis

Step-by-Step Pipeline:

1. **Data Preparation:** encoding all predictor variables into a model matrix and extracting the response variable representing class labels
2. **Compute Scatter Matrices:**
   - S_W: Covariance within classes.
   - S_B: Class mean deviations from overall mean.
3. **Solve:** $S_W^{-1} S_B v = \lambda v$
4. **Select Top Components:** Based on largest eigenvalues.
5. **Project and Classify:** New data points are projected onto these discriminant axes. Use class centroids in reduced space for classification.

.

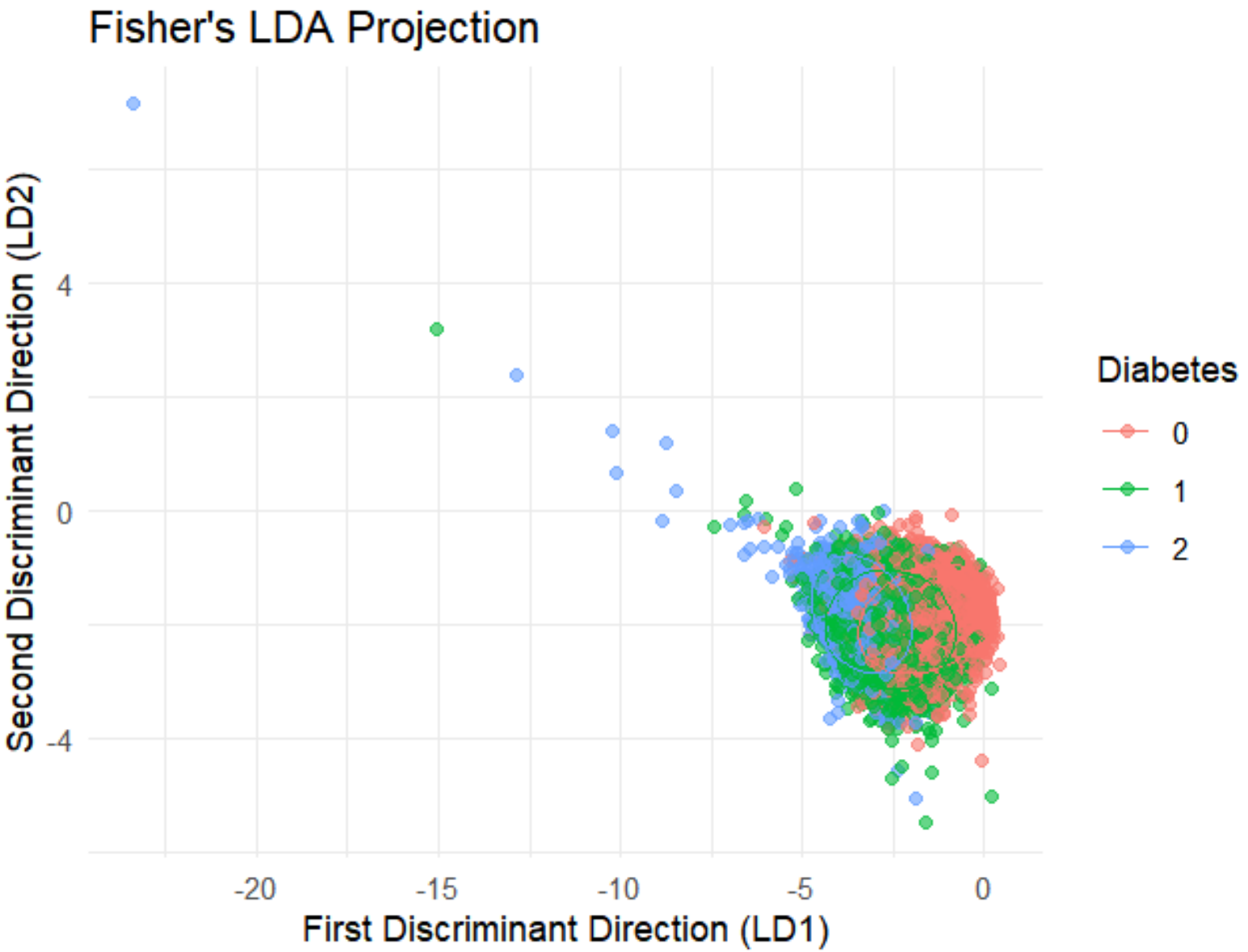
Projected Group Means

Table: Group Means in Linear Discriminant Space

Class	LD1	LD2	LD3
0	$-2.063632 + 0i$	$1.352984 + 0i$	$0.6417659 + 0i$
1	$-2.923893 + 0i$	$1.522407 + 0i$	$0.6417659 + 0i$
2	$-3.734283 + 0i$	$1.260310 + 0i$	$0.6417659 + 0i$

Fisher's Discriminant Projection

2D Visualization of Class Separation



Fisher Projection: Class Interpretation

What the 2D Space Tells Us

Insights:

- **LD1 (x-axis):** Strongly separates diabetic from non-diabetic groups.
- **LD2 (y-axis):** Is not able create much separation among the groups.
- **Diabetic vs Prediabetic:** Substantial overlap, reflecting biological ambiguity.
- **Outliers:** Negative LD1 extremes suggest atypical high-risk profiles.

3D Fisher Projection

Visualizing LD1, LD2, and LD3

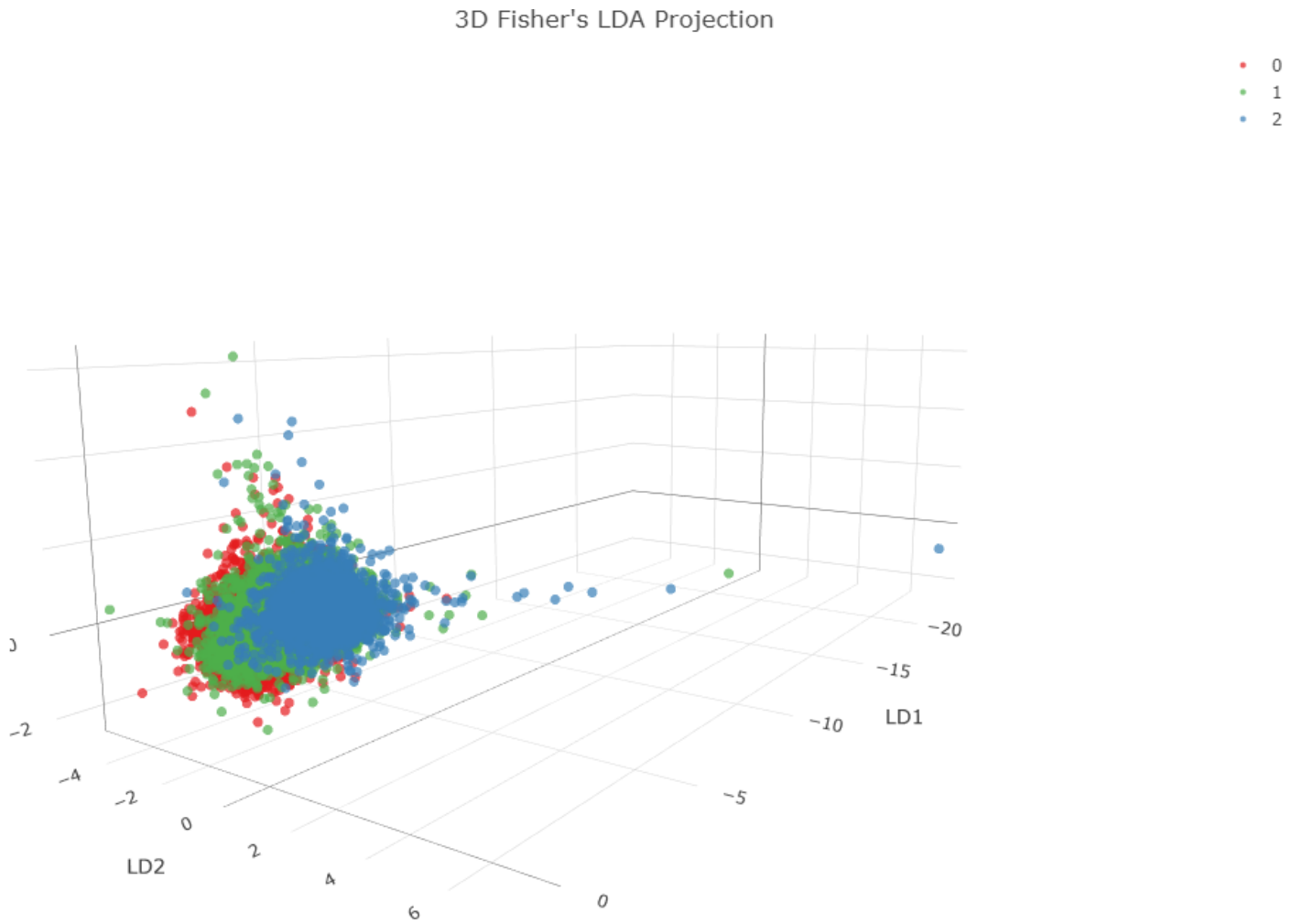


Figure: LD1 vs LD2 vs LD3: Enhanced separation with third discriminant axis.

Fisher Discriminants: Key Insights

Explained Variance and Clinical Interpretability

Class Distribution Patterns:

- Diabetic points tightly clustered along LD1.
- Normal and prediabetic overlap more in LD2-LD3 space.

Clinical Interpretation:

- LD1 – Strongly tied to metabolic syndrome indicators.
- LD2 – May reflect dietary or lifestyle variance.
- LD3 – Captures nuanced biochemical effects.

Table: Variance Explained by Discriminants

Component	Variance Explained
LD1	68.2%
LD2	21.7%
LD3	8.1%
Remaining	2.0%

Fisher LDA: Model Implementation Overview

1. **Data Prep:** Model matrix $\mathbf{X}$, labels $\mathbf{y}$.
2. **Fit LDA:** Class means $\boldsymbol{\mu}_k$, shared covariance.
3. **Compute Scatters:** $\mathbf{S}_W$, $\mathbf{S}_B$.
4. **Eigen Decomp:** $\mathbf{S}_W^{-1}\mathbf{S}_B \rightarrow$ LDs.
5. **Project:** Test data $\mathbf{Z}_{\text{test}} = \mathbf{X}_{\text{test}}\mathbf{V}_{1:2}$.
6. **Distance Calc:** Euclidean to class centroids $d_k(\mathbf{z})$.
7. **Classification:** Assign to closest class $\hat{y} =_k d_k(\mathbf{z})$.

Confusion Matrix and Model Performance

Three-Class Classifier Evaluation

Confusion Matrix

2*Predicted	Actual		
	0	1	2
0	2031	572	22
1	692	847	144
2	217	628	549

Class-Specific Metrics

Metric	Class 0	Class 1	Class 2
Sensitivity	0.691	0.414	0.768
Specificity	0.785	0.771	0.831
PPV	0.774	0.503	0.394
NPV	0.705	0.701	0.961

- **Accuracy**: 60.1% (CI: 58.8-61.4%)
- **Kappa**: 0.363 (Fair agreement)
- **Baseline**: 51.6% (No Information Rate)

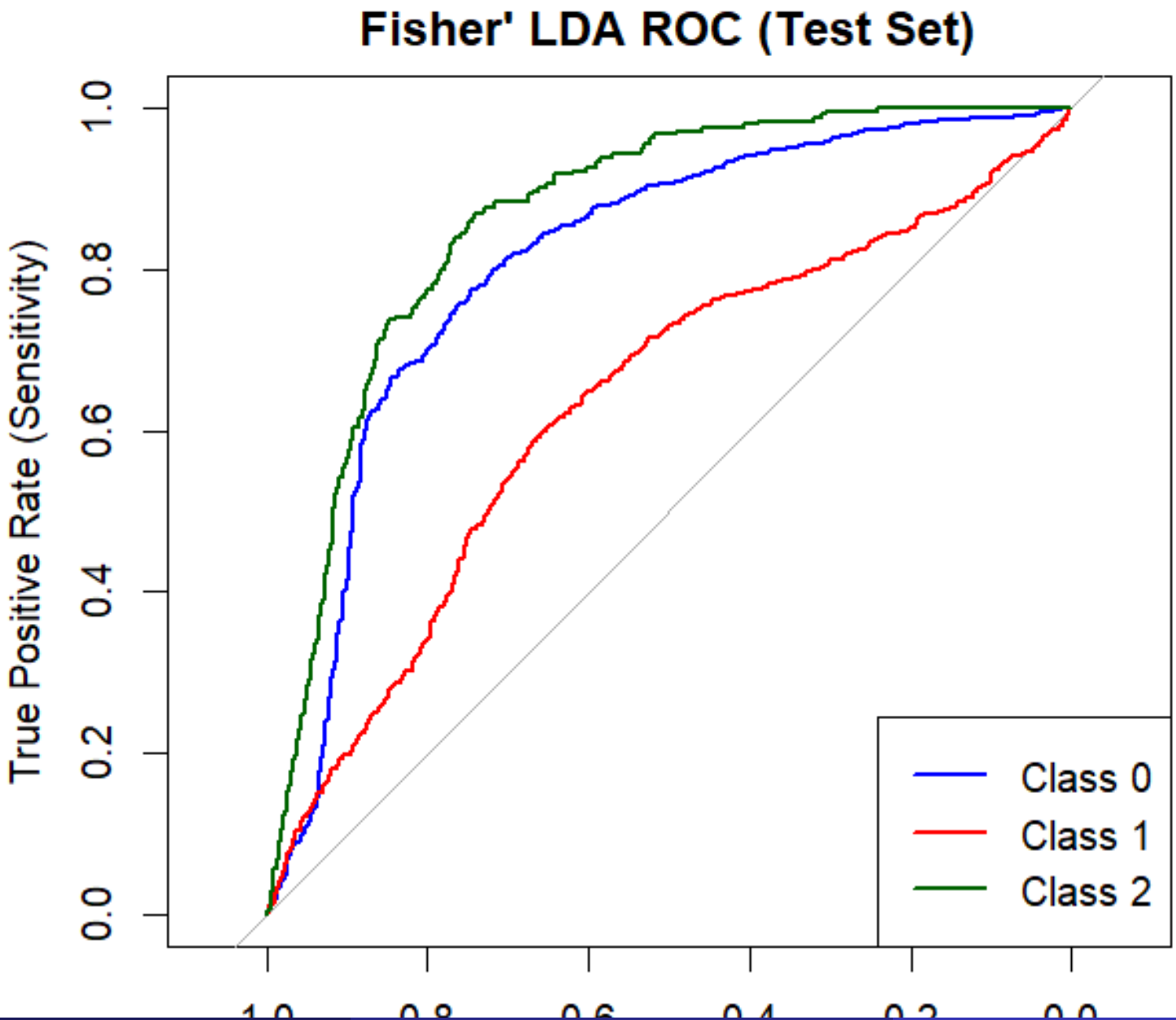
Key Observations

- Class 2: Highest sensitivity (76.8%)
- Class 1: Lowest performance (41.4% recall)
- Strong false positives between 0/1 classes

Key Observations

- **Class Imbalance:** Strong Class 2 performance (sensitivity = 0.768) despite low prevalence (12.5%)
- **Misclassification:** Class 1 has the lowest sensitivity (0.414); often confused with Class 0
- **Distance-based Discrimination:** Balanced accuracy:
  - Class 1: 0.593
  - Class 2: 0.799

ROC Curve Analysis



ROC Curve Analysis: Observations

Observations:

- **Class 2**: Highest AUC — strong diabetic detection
- **Class 1**: Weakest curve — poor prediabetes identification
- **Class 0**: Balanced sensitivity and specificity
- Stable ranking quality from linear boundaries

10-Fold Stratified Cross-Validation

1. **Data Partitioning:**
   - Stratified sampling ensures consistent class distributions.
   - Each fold used once for testing, others for training.
2. **Training Phase:**
   - Compute class means $\boldsymbol{\mu}_k$, global mean $\bar{\boldsymbol{\mu}}$.
   - Estimate $\mathbf{S}_W$, $\mathbf{S}_B$; solve for discriminant directions via:

$$\mathbf{S}_W^{-1}\mathbf{S}_B$$

Projection and Classification

Within Each Fold:

- **Projection:** Reduce to 2D discriminant space:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}_{1:2}$$

- **Centroid Computation:** $\boldsymbol{\mu}_k^{\mathbf{Z}}$
- **Classification:**

$$d_k(\mathbf{z}) = \|\mathbf{z} - \boldsymbol{\mu}_k^{\mathbf{Z}}\|_2 \quad \Rightarrow \quad \hat{y} =_k \; d_k(\mathbf{z})$$

Aggregated model performance

Confusion Matrix:

Prediction \ Reference	0	1	2
0	6977	1863	99
1	2351	2686	516
2	736	2012	1771

Overall Statistics:

- Accuracy: 0.6014
- 95% CI: (0.5944, 0.6084)
- No Info Rate: 0.5294
- P(Acc ¿ NIR): $< 2.2 \times 10^{-16}$
- Kappa: 0.3576
- McNemar's Test: $< 2.2 \times 10^{-16}$

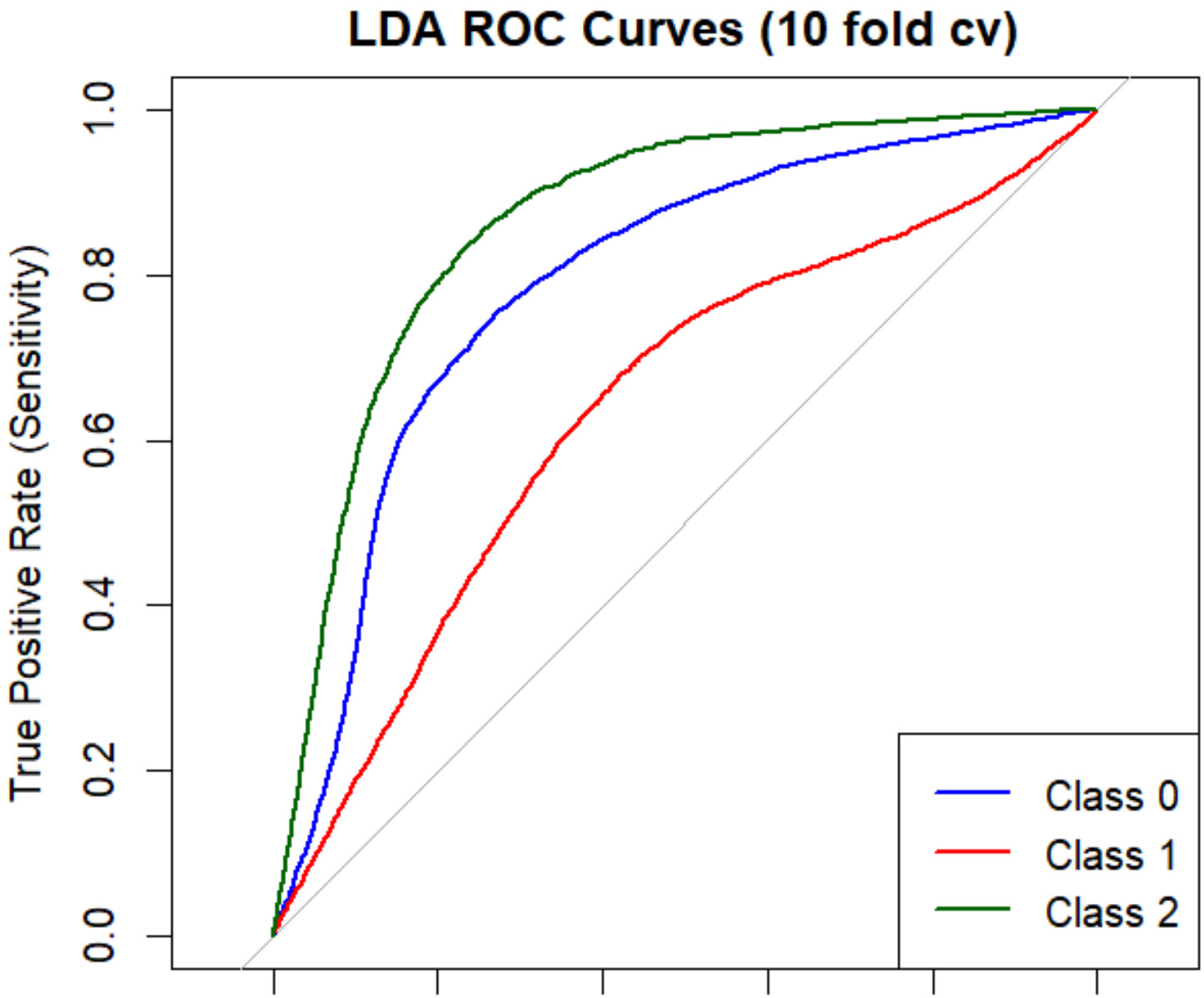
Per-Class Statistics:

Metric	Class 0	Class 1	Class 2
Sensitivity	0.6933	0.4094	0.7423
Specificity	0.7807	0.7697	0.8347
PPV	0.7805	0.4837	0.3919
NPV	0.6935	0.7121	0.9576
Prevalence	0.5294	0.3451	0.1255
Detection Rate	0.3670	0.1413	0.0932
Detection Prev.	0.4702	0.2921	0.2377
Balanced Acc.	0.7370	0.5896	0.7885

Key Observations

- **Consistency:** Maintains 60.1% accuracy despite class overlaps
- **Class 2:** Improved generalization with 1,771 correct predictions
- **Class 1:** High confusion with Class 0 (2,351 instances)
- **Linear Projection:** Preserves structure but limits non-linear separability

ROC Curve (10-Fold CV)



ROC Curve Insights

- **Class 2 (AUC = 0.862):** Strong separability
- **Class 0 (AUC = 0.787):** Balanced discrimination
- **Class 1 (AUC = 0.638):** Poorer separation, matches confusion patterns

Clinical Interpretations:

- **Class 2:** High AUC implies distinct diabetic markers
- **Class 1:** Lower AUC due to:
  - Overlap with Class 0
  - Internal heterogeneity
  - Potential need for new biomarkers

Limitations

- **Distance-based classification** Assumes uniform distributions
- **AUC bias:** May overestimate performance for smaller classes
- **Variability:** Fold-specific discriminant directions may differ

Conclusion

- **Discriminant Analysis:** A valuable tool for predicting and classifying diabetes.
- **Key Predictors:** Effective use of health indicators such as bp, BMI, and lipid profiles.
- **Assumption Handling:** Method remains robust with proper preprocessing and transformation, even when assumptions are slightly violated.
- **Clinical Impact:** Supports early detection of at-risk individuals, enabling timely intervention and improved management strategies.

"Better data-driven prediction leads to better prevention."

Thank You!

Any Questions?