Linear Models Graduate Project

Latherial Calbert and Natalie Huey

December 5th 2024, MATH 550

1 Introduction

For our Linear Models project, we wanted to find an established dataset from the textbook *A Modern Approach to Regression Using R* by Simon Sheather. In order to get the most out of the course, we decided to use a dataset that was introduced in the final chapter of the course, Logistic Regression. This way, we could use a culmination of the skills we learned throughout MATH 550 for this project. The HeartDisease data set caught our eye as we are both interested in the biomedical application of statistics. This data is used to predict whether a patient has heart disease based on five different predictors: systolic blood pressure (x_1), cholesterol (x_2), family history (x_3, yes=1), obesity (x_4), and age (x_5).

2 Mathematical Explanations of the Methods and Models

2.1 Logistic Regression Model

Logistic regression models the probability $P(y = 1 \mid \mathbf{x})$ for binary outcomes using the logistic function:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_5 x_5))}.$$

The log-odds (logit) is given by:

$$\text{logit}(P(y = 1 \mid \mathbf{x})) = \log\left(\frac{P(y = 1 \mid \mathbf{x})}{1 - P(y = 1 \mid \mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_5 x_5.$$

For predictors that are skewed, we apply transformations (like logarithms) to improve linearity in the logit:

$$\text{logit}(P(y = 1 \mid \mathbf{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 \log(x_1) + \beta_5 \log(x_2) + \beta_6 \log(x_4) + \beta_7 \log(x_5).$$

2.2 Assumptions of the Logistic Regression Model

The logistic regression model assumes:

1. **Binary Response Variable:** The response variable y is binary.
2. **Independent Observations:** Each observation is independent, with the likelihood given by:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(P(y_i = 1 \mid \mathbf{x}_i)\right)^{y_i} \left(1 - P(y_i = 1 \mid \mathbf{x}_i)\right)^{1-y_i}.$$

3. **Linearity in Logit:** The relationship between the predictors and the log-odds is linear, though transformations can be applied to achieve this linearity.

2.3 Gaussian Kernel Density Estimation (KDE) and Feature Transformation

For predictors that are not normally distributed (i.e., are skewed), we use Gaussian Kernel Density Estimation (KDE) to inspect the data distribution. For example, we may transform skewed predictors by applying logarithms (e.g., $\log(x_1)$) to make the distribution more normal and potentially improving linearity with the log-odds.

2.4 Model Fitting and Selection

The logistic regression model is fit by maximizing the log-likelihood:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log P(y_i = 1 \mid \mathbf{x}_i) + (1 - y_i) \log(1 - P(y_i = 1 \mid \mathbf{x}_i))\right].$$

Model selection is guided by the Akaike Information Criterion (AIC), which balances fit and complexity:

$$\text{AIC} = -2\ell(\boldsymbol{\beta}) + 2k,$$

where k is the number of parameters. Backward selection is used to minimize AIC and select the most parsimonious model.

2.5 Final Model

The final logistic regression model after applying transformations and selecting predictors is:

$$\text{logit}(P(y = 1 \mid \mathbf{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 \log(x_1) + \beta_5 \log(x_2) + \beta_6 \log(x_4) + \beta_7 \log(x_5),$$

with the following coefficients:

$$\begin{array}{llll} \beta_0 = 78.496, & \beta_1 = 0.106, & \beta_2 = 0.912, & \beta_3 = 0.442, \\ \beta_4 = -14.808, & \beta_5 = 1.095, & \beta_6 = -13.284, & \beta_7 = 2.285. \end{array}$$

The residual deviance is 482.3 with 454 degrees of freedom, and the AIC is 498.3.

2.6 Model Evaluation

To evaluate the model, we use marginal model plots (MMPs). These plots compare the observed values with the fitted values for each predictor:

1. **Model Line:** The fitted values from the logistic regression model.
2. **LOESS Line:** A smoothed curve (LOESS) representing the relationship between the predictor and the response.

The goal is for the model line and the LOESS line to align. If they both show a linear relationship or follow the same trend, this suggests the model is appropriately specified.

2.7 Interpretation of Coefficients

For example, the odds ratio for predictor x_3 is:

$$\text{Odds Ratio for } x_3 = \exp(\beta_2) = \exp(0.912) \approx 2.49,$$

indicating that the odds of the outcome (e.g., heart disease) increase by approximately 149% for a unit increase in x_3.

2.8 Conclusion

The final logistic regression model, with appropriate transformations and selected predictors, is robust and provides valuable insights into the relationship between predictors and the log-odds of the response. The model has been evaluated using MMPs, and the coefficients have been interpreted to understand the effects of each predictor on the outcome.

3 Data Analysis

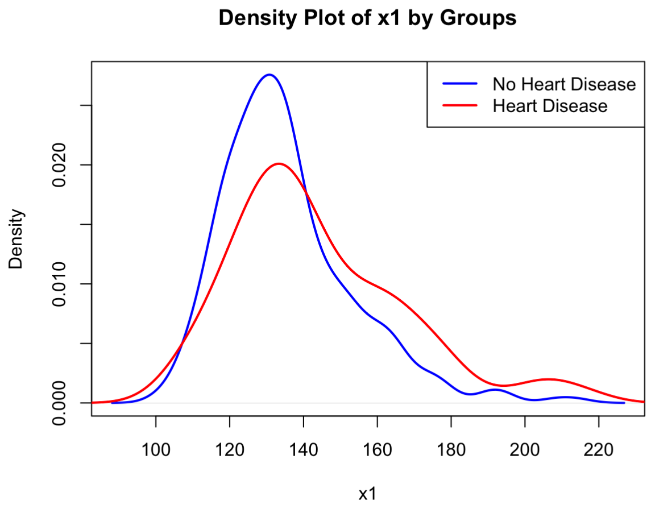
First, we created the logistic model, which gives us the following output:

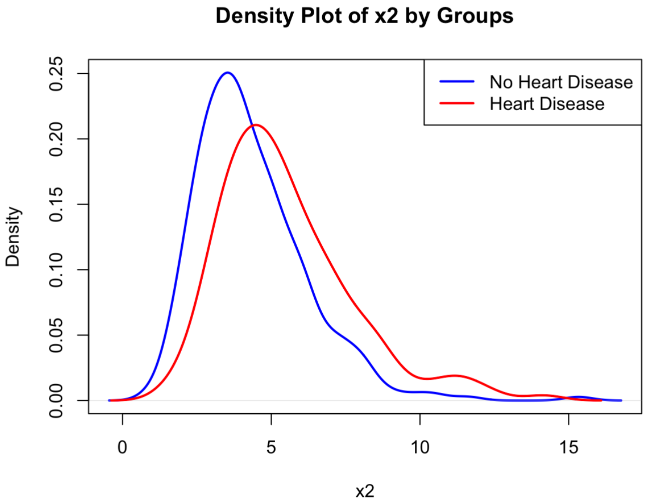
```
    Call:
glm(formula = HeartDisease ~ x1 + x2 + x3 + x4 + x5, family = binomial,
    data = HeartDisease)

Coefficients:
```

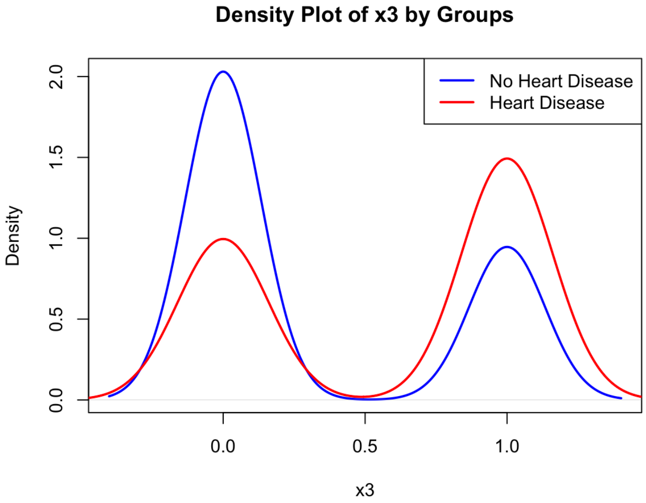
```
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.313426   0.943928  -4.570 4.89e-06 ***
x1           0.006435   0.005503   1.169 0.242227    
x2           0.186163   0.056325   3.305 0.000949 ***
x3           0.903863   0.221009   4.090 4.32e-05 ***
x4          -0.035640   0.028833  -1.236 0.216433    
x5           0.052780   0.009512   5.549 2.88e-08 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1 
1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 493.62  on 456  degrees of freedom
AIC: 505.62

Number of Fisher Scoring iterations: 4
```

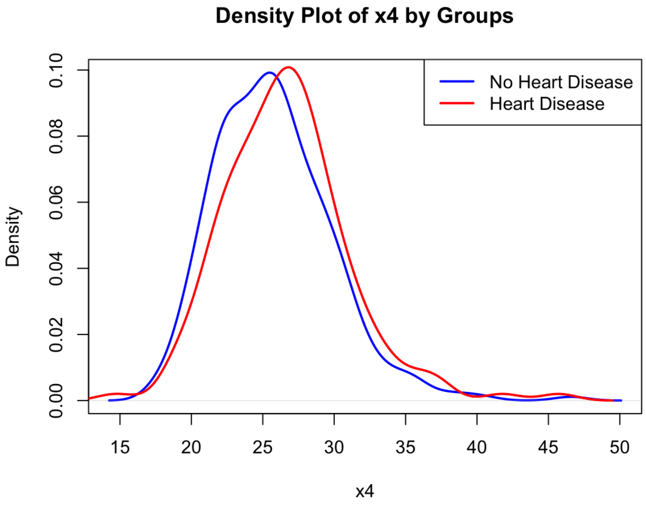
In order to ensure that a logistic regression model should be used for our data set, we run a hypothesis test and find a p-value of 0.1084517, meaning the logistic model is viable.

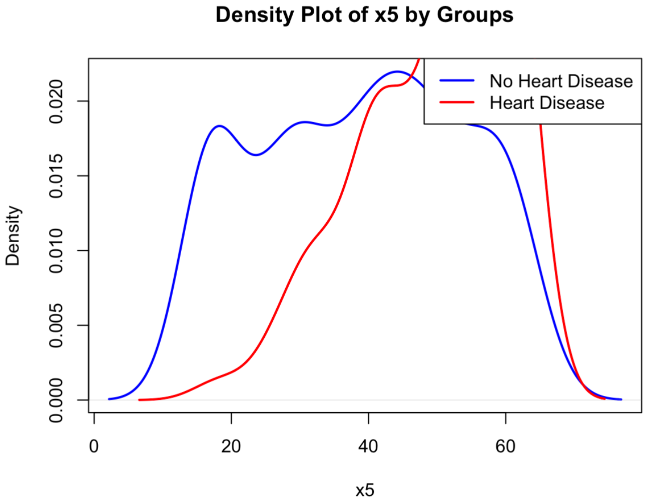
To assess the assumption of linearity between the log-odds (logit) and each predictor in the logistic regression model, we first examine the Gaussian Kernel Density Estimation plots. We look for plots that are asymmetric, skewed, or not normal. Below are our plots:



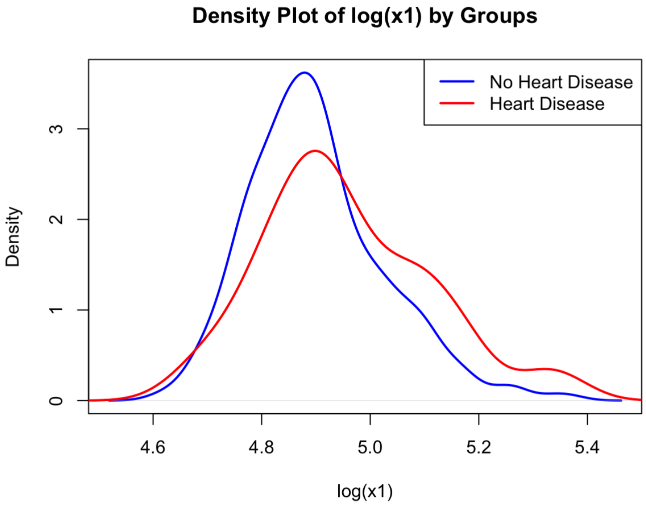

Density Plot of x2 by Groups
No Heart Disease
Heart Disease
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
x2

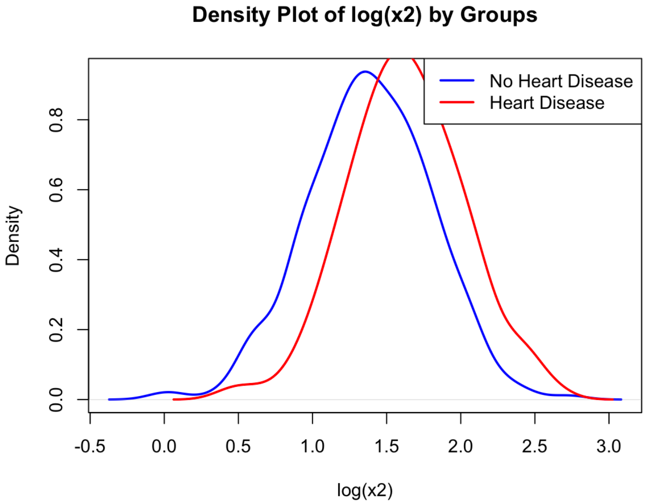

Density Plot of x3 by Groups
No Heart Disease
Heart Disease
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
x3

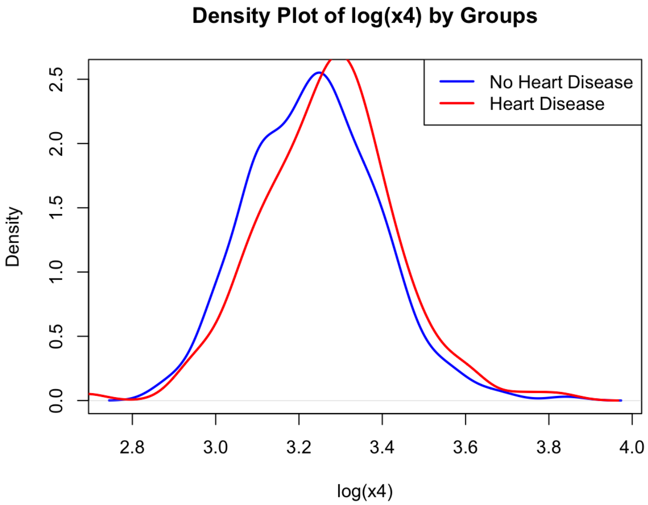


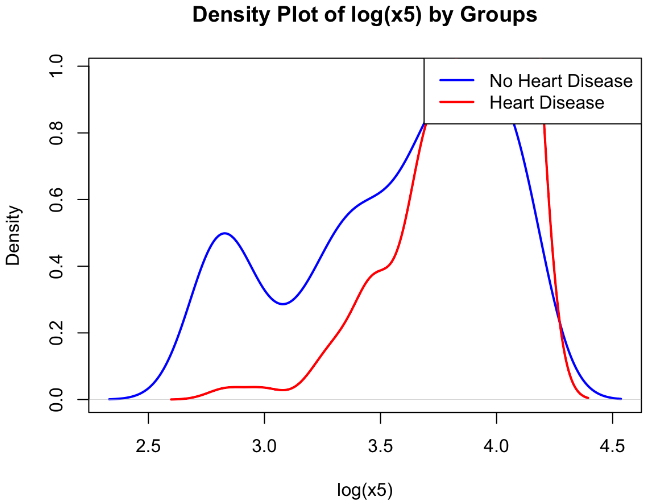


As one can see, none of these predictors are particularly normal, so we will add logarithm terms for all, with the exception of x_3 because it is binary. We examined the density plots for the logarithm terms to ensure that these terms are less skewed and more normal.


Density Plot of log(x1) by Groups
No Heart Disease
Heart Disease
Density
0
1
2
3
4.6
4.8
5.0
5.2
5.4
log(x1)


Density Plot of log(x2) by Groups
No Heart Disease
Heart Disease
Density
0.0
0.2
0.4
0.6
0.8
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(x2)



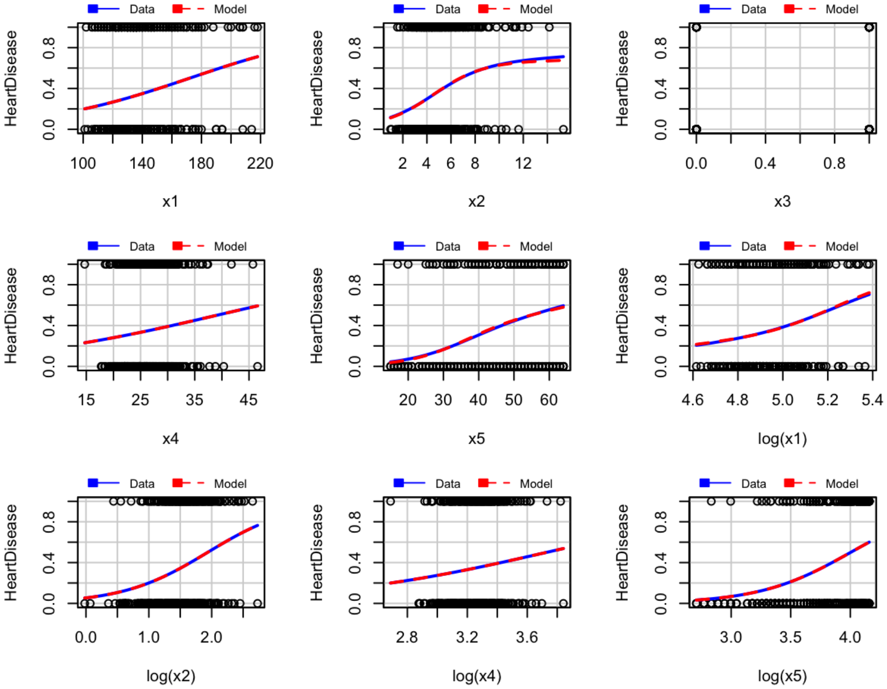


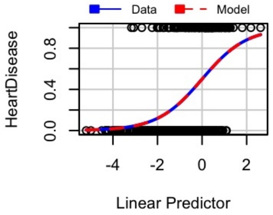
Below is the output from the full logistic model with the transformed predictors added:

```
    Call:
glm(formula = HeartDisease ~ x1 + x2 + x3 + x4 + x5 + log(x1) +
    log(x2) + log(x4) + log(x5), family = binomial, data = HeartDisease)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 76.85660   34.81695   2.207   0.0273 *
x1           0.10854    0.05343   2.031   0.0422 *
x2           0.02020    0.19163   0.105   0.9161
```

```
x3            0.90670    0.22470   4.035 5.46e-05 ***
x4            0.47043    0.22452   2.095   0.0361 *
x5           -0.04667    0.05319  -0.877   0.3802
log(x1)     -15.06600    7.89649  -1.908   0.0564 .
log(x2)       0.98637    1.03403   0.954   0.3401
log(x4)     -14.18659    6.18588  -2.293   0.0218 *
log(x5)       4.13061    2.16508   1.908   0.0564 .
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1
1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 481.49  on 452  degrees of freedom
AIC: 501.49

Number of Fisher Scoring iterations: 5
```

Next we use Marginal Model Plots to further visualize our data:





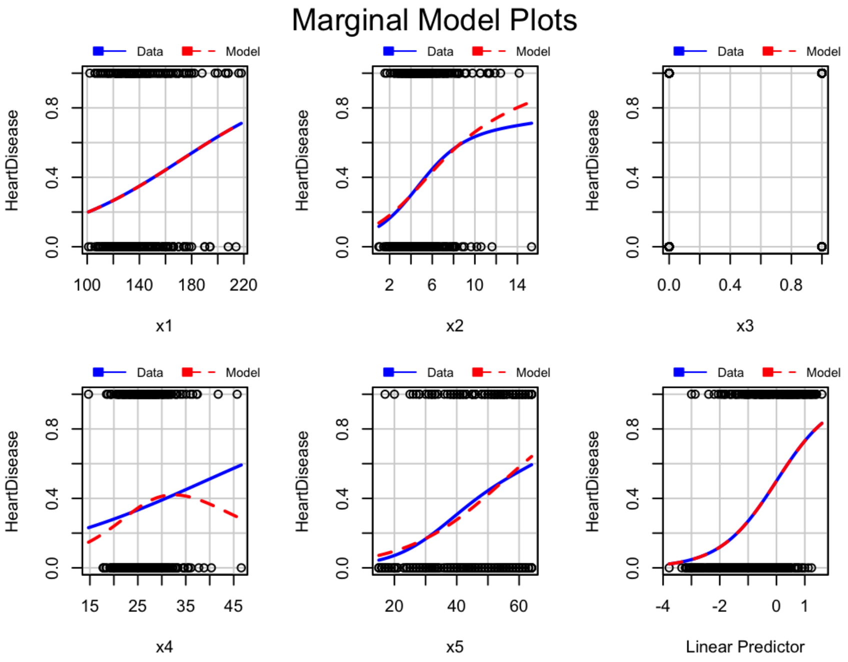
The final model is further optimized by minimizing its complexity, guided by the Akaike Information Criterion (AIC), which balances model fit and parsimony.

This tells us that our final model and summary should be:

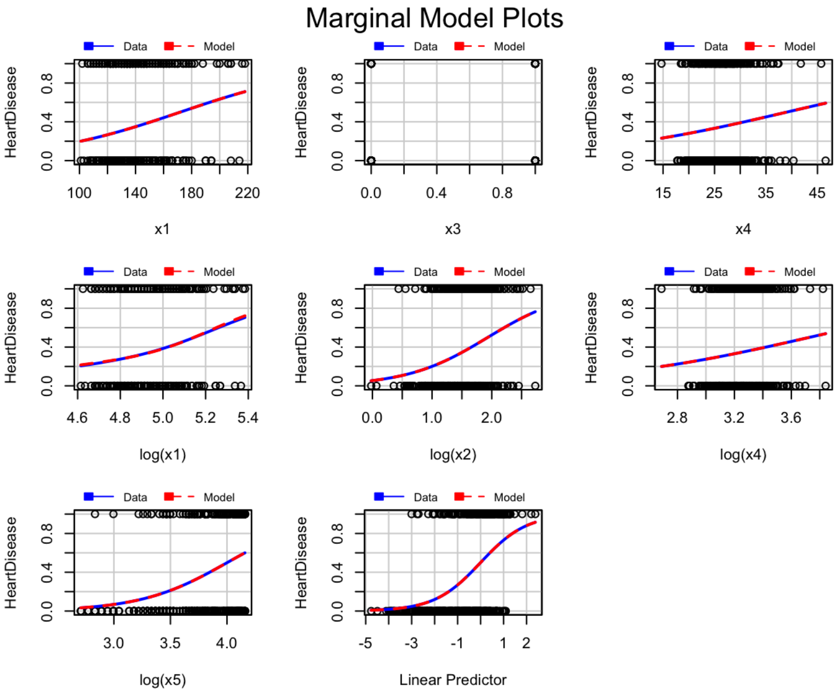
```
    Call:
glm(formula = HeartDisease ~ x1 + x3 + x4 + log(x1) + log(x2) +
    log(x4) + log(x5), family = binomial, data = HeartDisease)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  78.49626   34.42745   2.280 0.022605 *
x1            0.10623    0.05306   2.002 0.045266 *
x3            0.91197    0.22475   4.058 4.96e-05 ***
x4            0.44153    0.21786   2.027 0.042699 *
log(x1)     -14.80803    7.84355  -1.888 0.059036 .
log(x2)       1.09449    0.30539   3.584 0.000338 ***
log(x4)     -13.28400    5.98725  -2.219 0.026506 *
log(x5)       2.28470    0.40224   5.680 1.35e-08 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 482.30  on 454  degrees of freedom
AIC: 498.3

Number of Fisher Scoring iterations: 5
```

Finally, we check the Marginal Model Plots for our final model in comparison with the original, to see that we have improved the model through these various steps.

Original:



Reduced and Transformed:



We can easily see that our new model fits the data better than the original, which is specifically evident when looking at the x_2, x_4, and x_5 plots in comparison with the logarithm plots.

This process of reducing features and optimizing normality ensures the robustness and efficiency of the logistic regression model.

4 Criticisms

4.1 Pros

- **Appropriate for Binary Response:** Logistic regression is ideal for binary outcomes, such as predicting the presence or absence of heart disease.
- **Improved Linearity with Transformations:** Since our predictors were skewed and not normal, the added logarithm terms improved linearity.
- **Significant Predictors Retained:** Statistically significant variables like cholesterol, family history, and obesity are kept, enhancing predictive power. We also included the logarithm terms for blood pressure and age, so all predictors are still included in the final model.
- **Adequate Sample Size:** The large sample size (462 observations) reduces overfitting risk and provides statistical power.
- **Model Optimization with AIC:** The model is optimized using the Akaike Information Criterion (AIC), balancing fit and simplicity.

4.2 Cons

- **Exclusion of Interaction Effects:** Excluding interaction terms between x_3 which was our binomial predictor of family history can ignore important relationships with other predictors.
- **Interpretation Complexity:** Logarithm transformations complicate the interpretation of coefficients, especially for transformed variables.
- **Model Complexity:** Despite simplifications, the final model remains complex, making it difficult to communicate to non-technical audiences.
- **Non-Linear Relationships:** While transformations address many nonlinearities, some may still persist, limiting model accuracy. In particular, the x_5 predictor is still not normal after we used the transformation which could be limiting.
- **Limited Generalizability:** The model is based on heart disease data, limiting its applicability to other populations.

- **Assumptions Dependence:** The model relies on assumptions like independence and linearity, which, if violated, could degrade performance.
- **Complexity in Refinement:** Continuous model adjustments (transformations, variable selection, interaction testing) can become time-consuming.

5 R Code

```
library(readr)
library(car)
HeartDisease <- read_csv("Desktop/HeartDisease.csv")
View(HeartDisease)
attach(HeartDisease)
original<-glm(HeartDisease~x1+x2+x3+x4+x5, data=HeartDisease,
    family = binomial)
summary(original)
out<-summary(original)
obs<-out$deviance
df<-out$df.residual
pvalue<-1-pchisq(obs,df)
mmps(original,layout=c(2,3))
# List of variable names to process
variables <- c("x1", "x2", "x3", "x4", "x5")

# Loop through each variable
for (var in variables) {
  # Extract the current variable
  current_var <- HeartDisease[[var]]

  # Basic density plot of the variable
  plot(density(current_var),
       main = paste("Density␣Plot␣of", var),
       xlab = var,
       col = "black",
       lwd = 2)

  # To compare groups: HeartDisease == 0 and HeartDisease == 1
  plot(density(current_var[HeartDisease$HeartDisease == 0]),
       col = "blue",
       main = paste("Density␣Plot␣of", var, "by␣Groups"),
       xlab = var,
       lwd = 2)
  lines(density(current_var[HeartDisease$HeartDisease == 1]),
        col = "red",
        lwd = 2)

  # Add a legend
  legend("topright",
         legend = c("No␣Heart␣Disease", "Heart␣Disease"),
         col = c("blue", "red"),
         lwd = 2)
}

# List of variable names to process
variables <- c("x1", "x2", "x3", "x4", "x5")

# Loop through each variable
for (var in variables) {
  # Log-transform the current variable
  log_var <- log(HeartDisease[[var]])

  # Basic density plot of the log-transformed variable
  plot(density(log_var),
       main = paste("Density␣Plot␣of␣log(", var, ")", sep = ""),
       xlab = paste("log(", var, ")", sep = ""),
       col = "black",
       lwd = 2)

  # To compare groups: HeartDisease == 0 and HeartDisease == 1
  plot(density(log_var[HeartDisease$HeartDisease == 0]),
       col = "blue",
       main = paste("Density␣Plot␣of␣log(", var, ")␣by␣Groups", sep = ""),
       xlab = paste("log(", var, ")", sep = ""),
       lwd = 2)
  lines(density(log_var[HeartDisease$HeartDisease == 1]),
        col = "red",
        lwd = 2)

  # Add a legend
  legend("topright",
         legend = c("No␣Heart␣Disease", "Heart␣Disease"),
         col = c("blue", "red"),
         lwd = 2)
}
logmodel<-glm(HeartDisease~x1+x2+x3+x4+x5+log(x1)+log(x2)+log(x4)+log(x5),
    data=HeartDisease, family = binomial)
mmps(logmodel,layout=c(2,3))
summary(logmodel)
null<-glm(HeartDisease~0, data=HeartDisease, family="binomial")
reg1b<-step(logmodel,scope=list(lower=null,upper=logmodel),
    direction ="backward")
final<-glm(HeartDisease~x1+x3+x4+log(x1)+log(x2)+log(x4)+log(x5),
    data=HeartDisease, family=binomial)
summary(final)
mmps(final,layout=c(3,3))
```