# Predicting Heart Failure from Risk Factor Data

Paige Galvan, Neha Deshpande, & Witlie Leslie

2024-11-25

## Introduction (Witlie)

For our project, we aimed to create a model that can predict the event of death as a result of heart failure from clinical patient data.

- Binary response variable "DEATH_EVENT", value of 1 indicating patient deceased (0 otherwise)
- 12 risk factor variables
  - 5 binary: anemia status, diabetes status, high blood pressure status, sex, and smoking status
  - 7 numerical: age, creatine phosphokinase level, ejection fraction, platelet concentration, serum creatine level, serum sodium level, and length of follow-up period

# Motivation (Neha)

- Investigating heart failure, a condition impacting millions worldwide, and exploring its complex causes and contributing factors.
- Spotting patterns between risk factors and how heart failure progresses to get a better understanding.

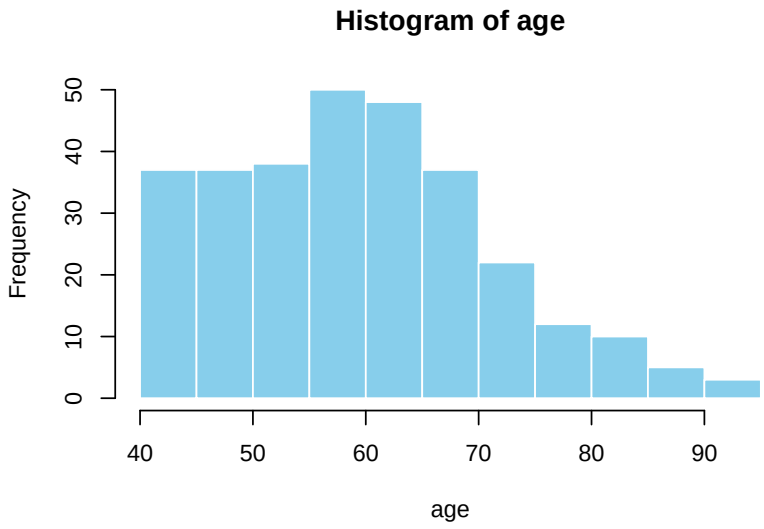
# Linearity and Normality (Neha)

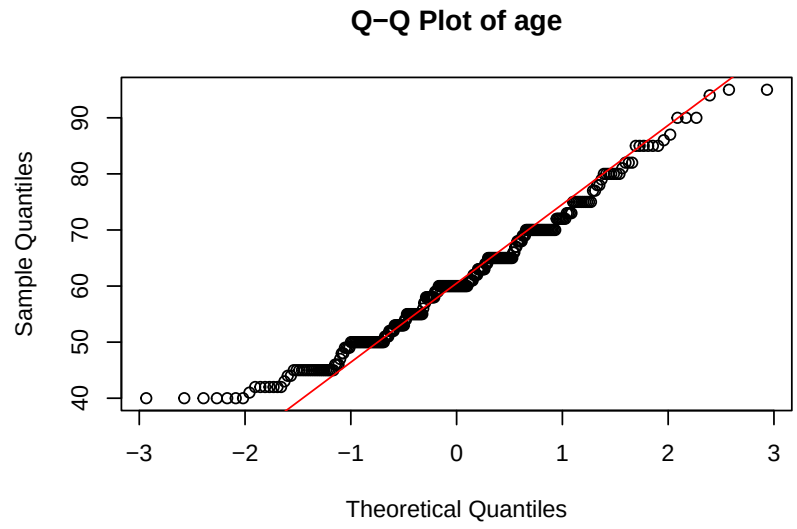
Normality

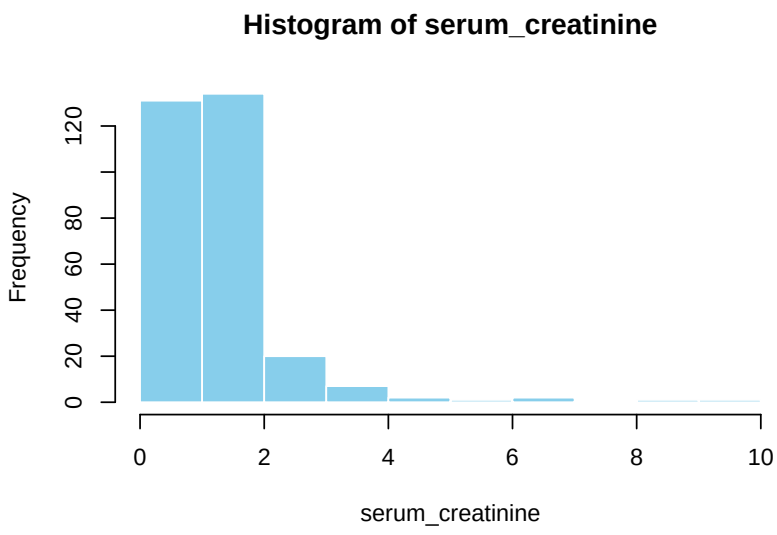
- Objective: Assess whether continuous predictor variables are normally distributed.
- Variables analyzed: Age, Creatinine Phosphokinase, Ejection Fraction, Platelets, Serum Creatinine, Serum Sodium.
- Testing residuals ensures the model fits the data well and detects patterns that might indicate a need for model improvement.
- Most variables were either right or left skewed, showing low normality, except for platelets and serum sodium, which were closer to normal
- Tests performed:

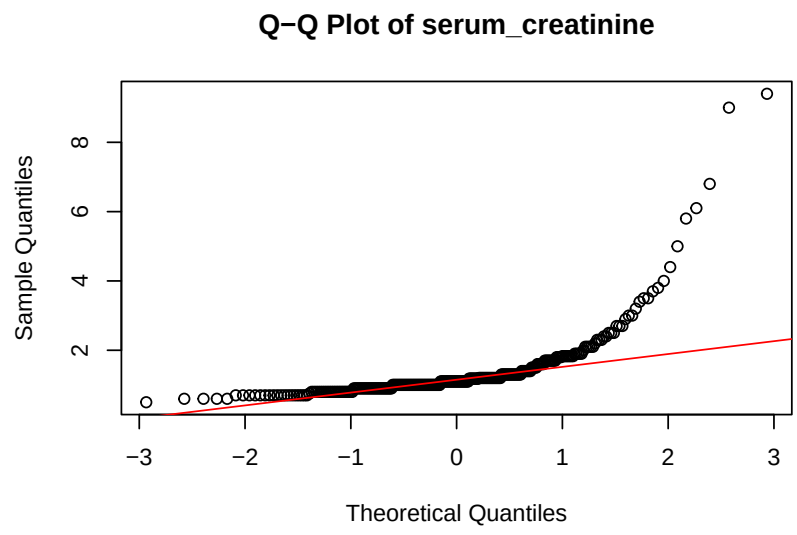
Shapiro-Wilk Test: p-values $< 0.05$ for all variables $\rightarrow$ None follow a normal distribution.

# Graphs


Histogram of age
Frequency
0
10
20
30
40
50
40
50
60
70
80
90
age


Q−Q Plot of age
Sample Quantiles
40
50
60
70
80
90
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histogram of serum_creatinine
Frequency
0
20
40
60
80
120
0
2
4
6
8
10
serum_creatinine

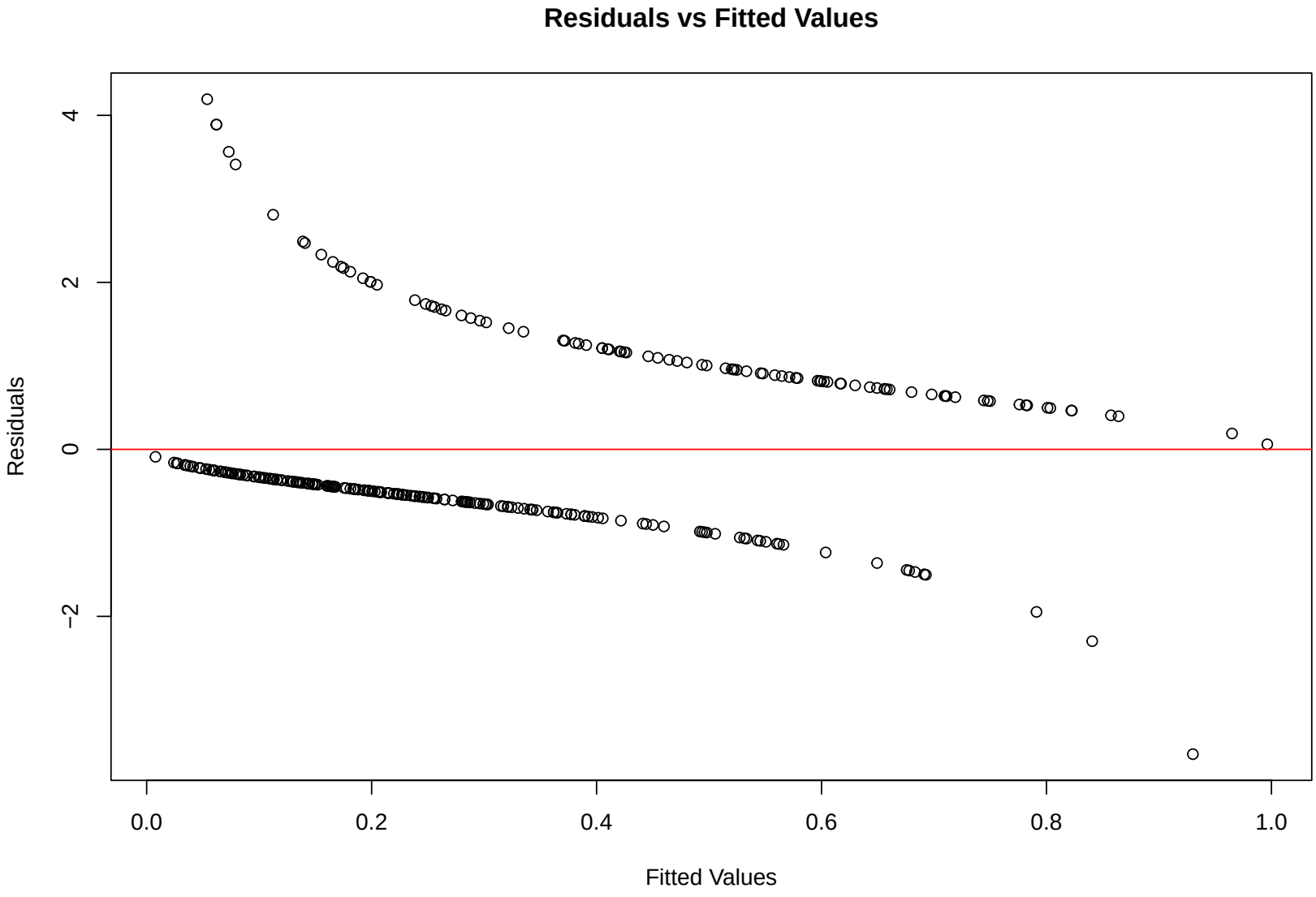

Q−Q Plot of serum_creatinine
Sample Quantiles
2
4
6
8
−3
−2
−1
0
1
2
3
Theoretical Quantiles

# Linearity

- Linearity assumes a straight-line relationship between the predictor variables and the log odds of the outcome in logistic regression
- Fitted linear trendlines show non linear relationships for most variables.
- Logistic regression results:

The residuals exhibit a curved pattern, indicating that the relationship between the predictors and the response variable may not be linear.

# Graphs


Residuals vs Fitted Values
Residuals
Fitted Values
4
2
0
−2
0.0
0.2
0.4
0.6
0.8
1.0

# Polynomial Model to Test Linearity

- Polynomial models allow us to model nonlinear relationships between predictors and the outcome
- The polynomial model identified several significant predictors of the likelihood of a death event, including age, serum creatinine, ejection fraction, and time. However, many other predictors, such as platelets, serum sodium, and factors like anaemia, diabetes, and smoking, were found to be statistically insignificant

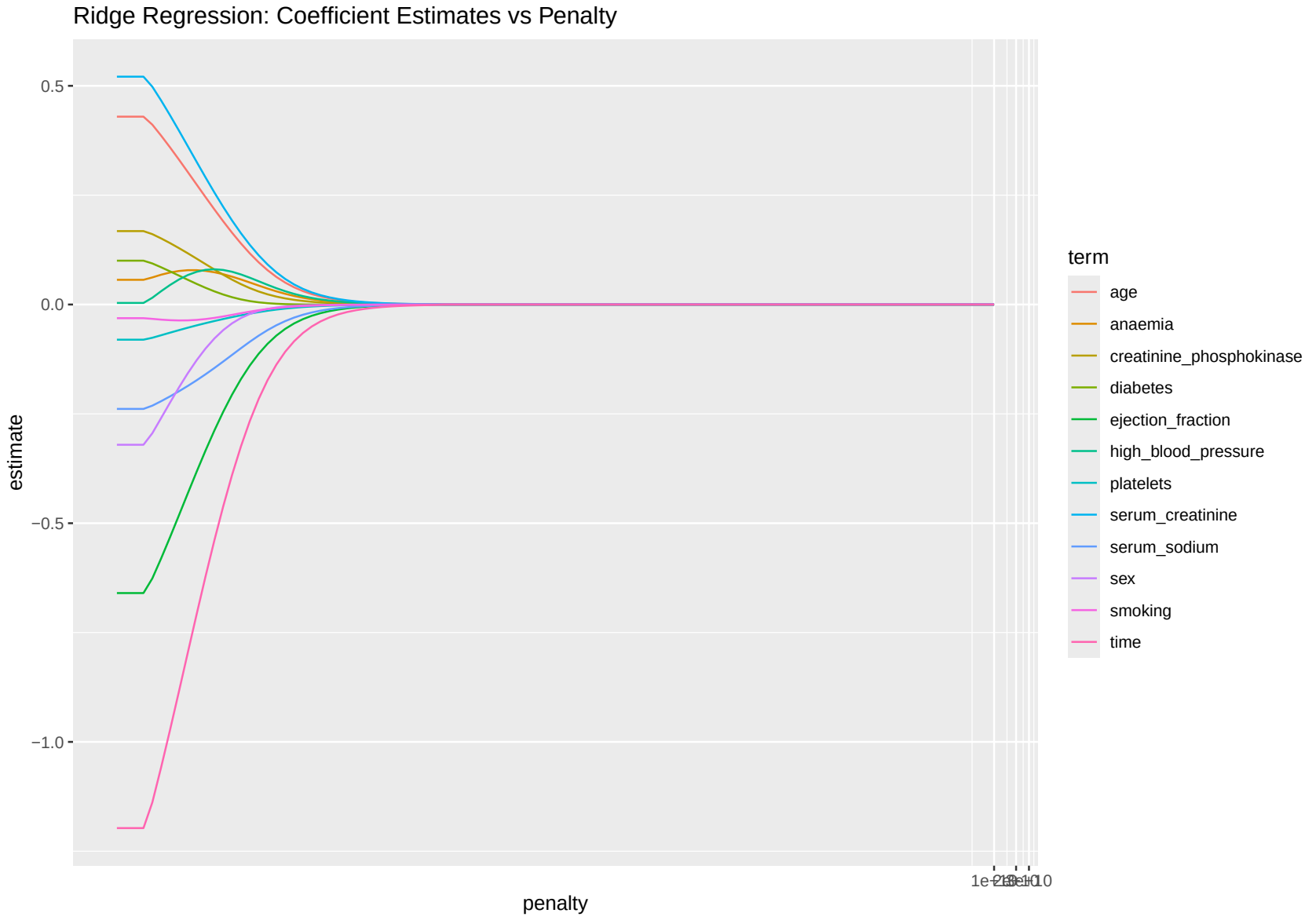
# SVM & Random Forest (Paige)

# Logistic Regression (Witlie)

Logistic regression was an obvious first choice in tackling this binary classification problem. I created 3 models to assess predictive performance with different regularization methods.

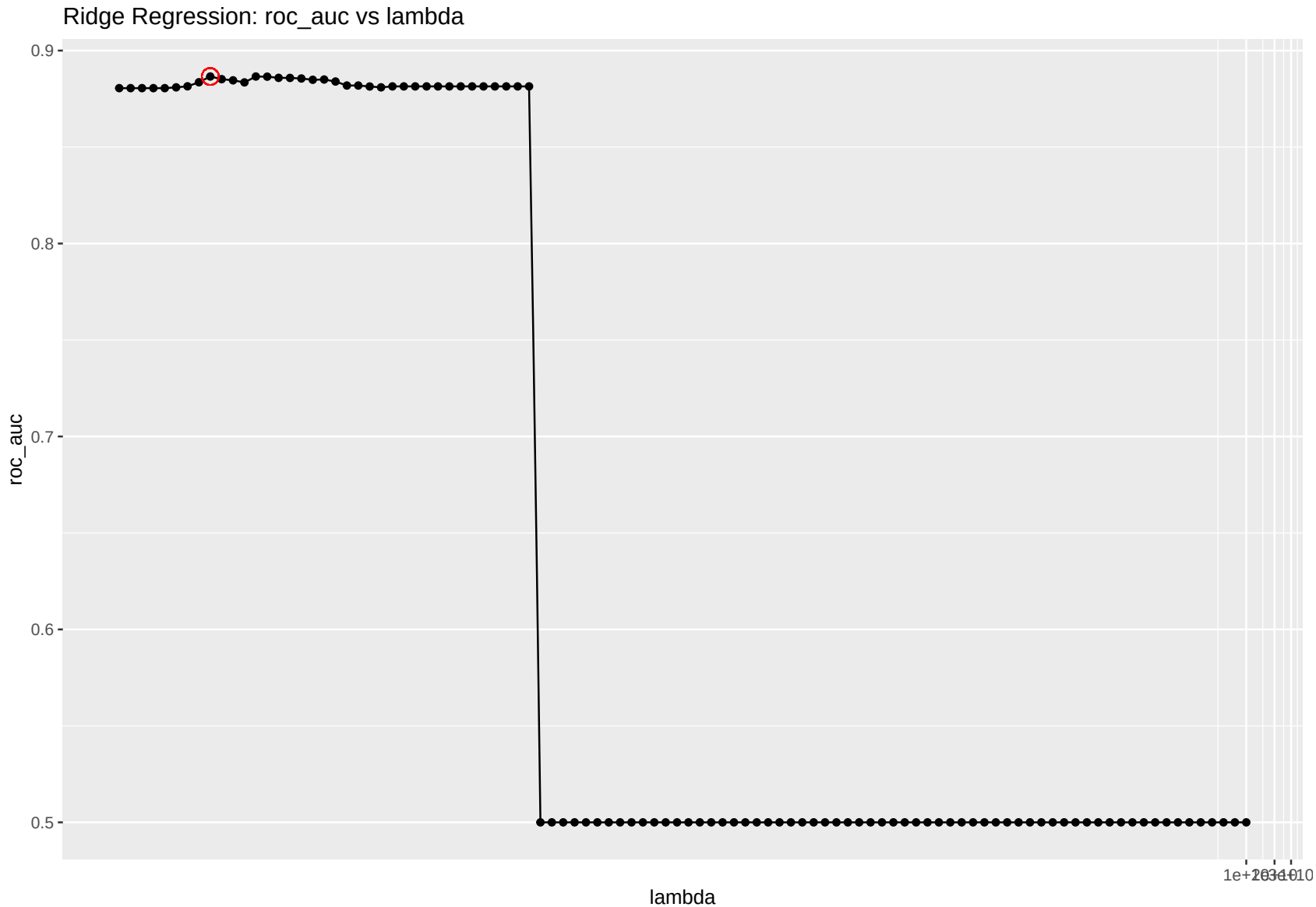
- Ridge Regression
- Lasso
- No regularization

# Ridge Regression



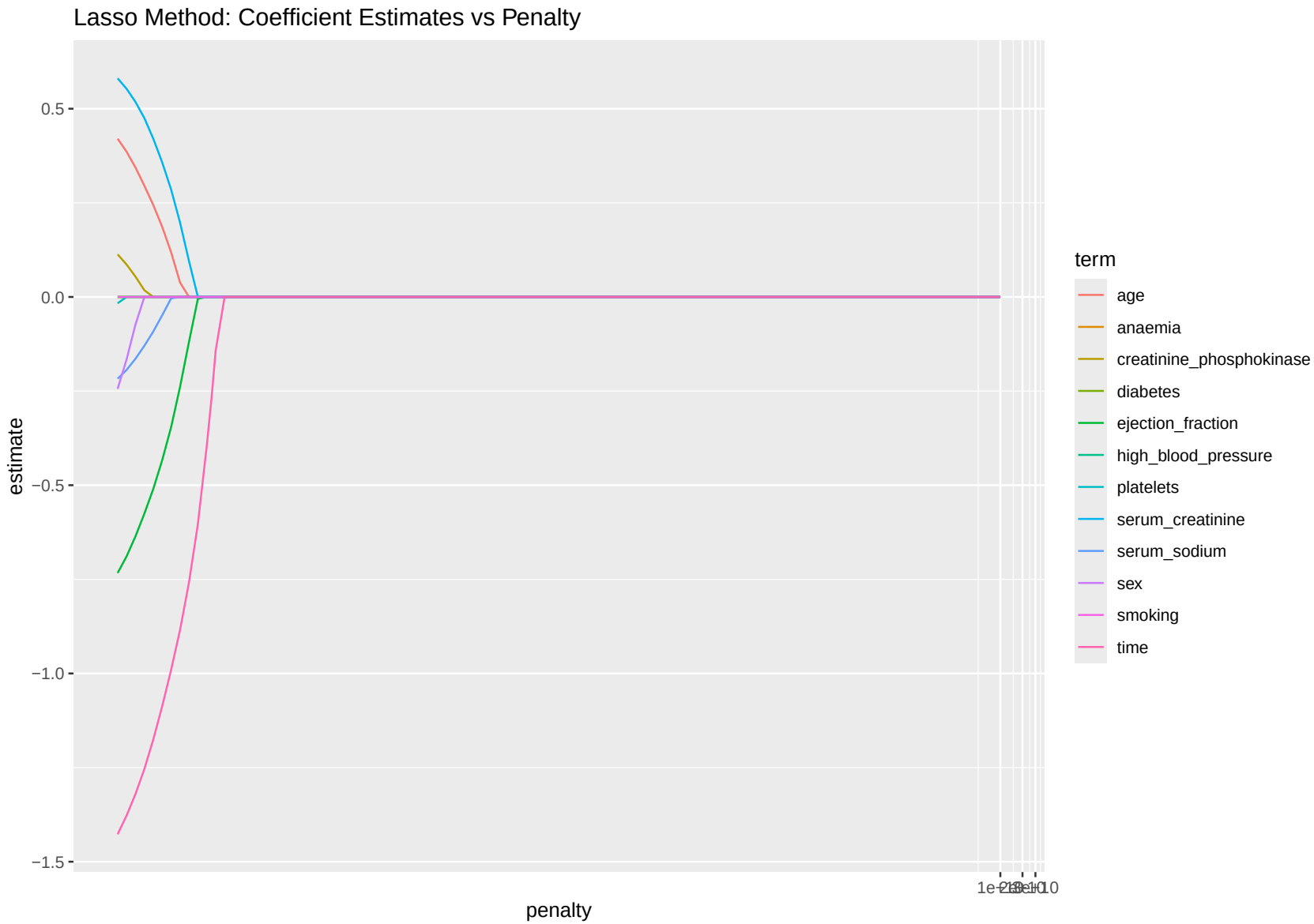
# Ridge Regression

Using 10-fold cross-validation, I found the lambda with the highest ROC-AUC value of 0.887 was lambda = 0.0933



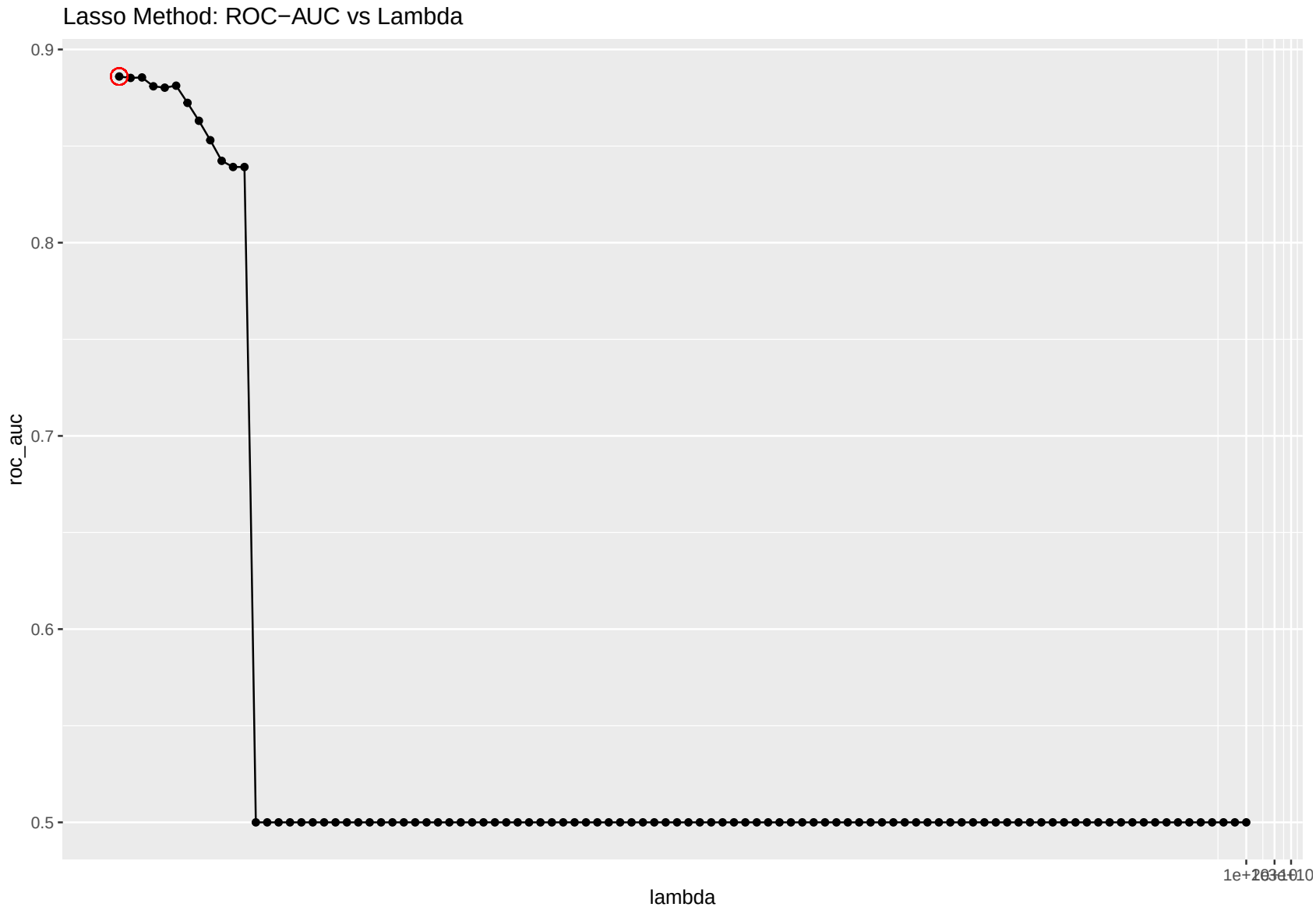
# Lasso

Lasso shows a steeper dropoff of coefficient estimates as a result of feature reduction



# Lasso

Using 10-fold cross validation, I found the lambda with the highest ROC-AUC value of 0.886 was lambda = 0.01



# Results

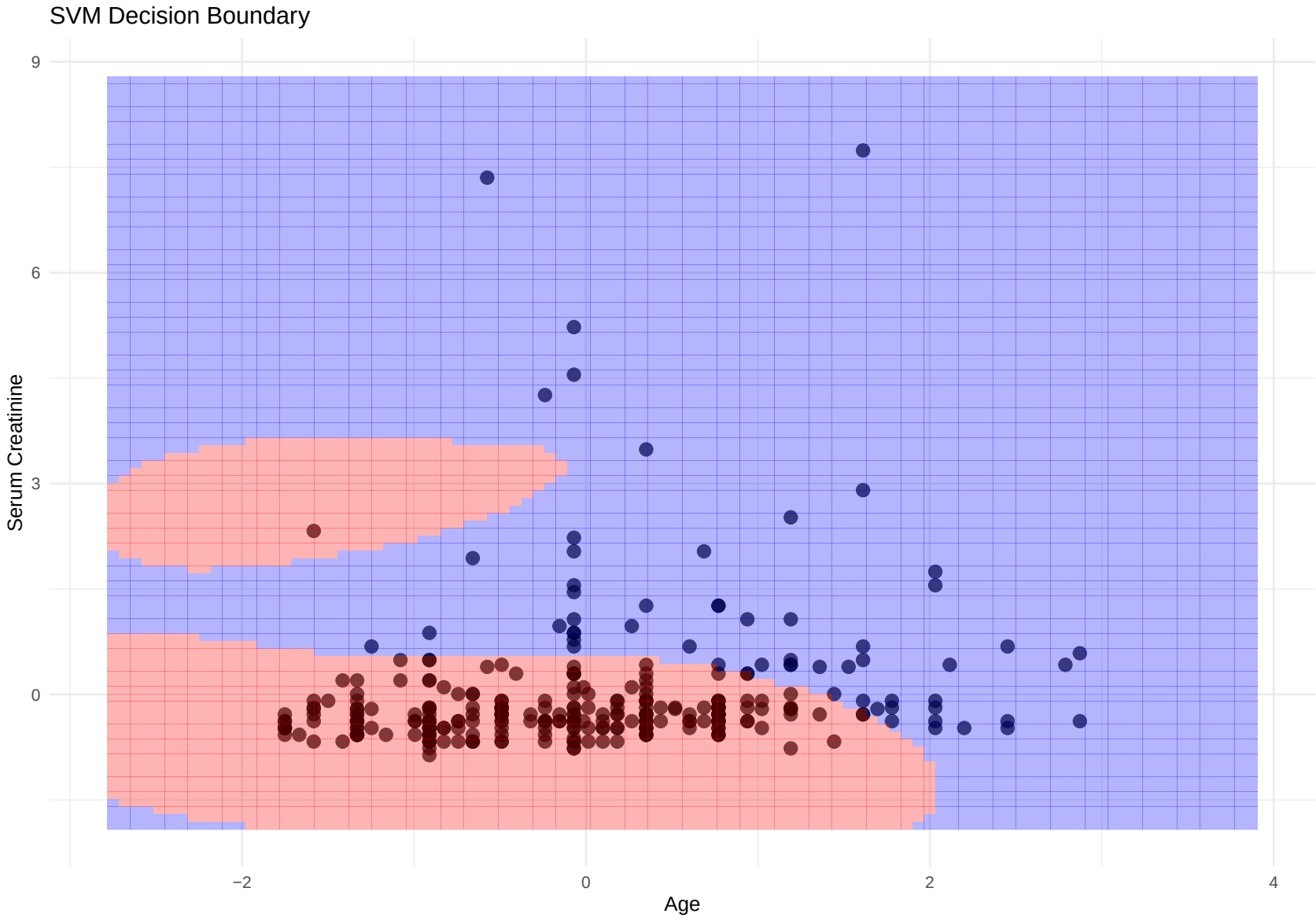
- No Regularization:
- Ridge Regression:
- Lasso:

# SVM (Paige)

- Predicting Death Event
- Like our Linear Regression methods we began by splitting the data into training and test data.
- I began by assessing the support variables and determining the acurracy of the model.

```
##
## Call:
## svm(formula = DEATH_EVENT ~ ., data = train_data, kernel
##     cost = 1, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  148
##
```

# SVM Model




- The SVM Boundry demonstrates that

# Random Forest (paige)

# Results (Paige)

# References

Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020). https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5