Compulsory Exercise 2: Age Prediction for Heart Failure
Group 20

Karianne Strand Bergem   Marte Ragnhild Hotvedt   Erlend Henriksen Winje

07 April, 2025

Abstract

This project explores a regression task using the Heart Failure Prediction dataset, published by Fedesoriano on Kaggle. While the dataset is often used for heart disease classification, we use it to estimate a person's age based on clinical and physiological measurements, treating age as the response variable and all other features as predictors. We apply and compare two statistical learning methods: multiple linear regression (MLR) and ridge regression. The goal is to assess how regularization in ridge affects prediction accuracy and model complexity, compared to the more interpretable baseline provided by MLR. We split the dataset into a 70% training set and a 30% test set to evaluate model performance. Prior to modeling, we conduct descriptive analysis to examine distributions, correlations, and variable relevance. Some features, such as maximum heart rate and exercise-induced angina, show moderate association with age, but no single variable is strongly predictive. Model assumptions for MLR are checked using diagnostic plots. Both models are evaluated using Mean Squared Error (MSE) and the coefficient of determination (R^2) on the test set. Results show that neither model achieves high predictive accuracy, with both yielding a root MSE around 8 years and R^2 below 0.3. Ridge regression generalizes slightly better due to its handling of multicollinearity. These results suggest that, while age can be partially predicted from the available features, accurate estimations are challenging with this dataset.

Introduction

In this project, we aim to predict a person's age using clinical and physiological variables from the Heart Failure Prediction dataset, originally published on Kaggle by Fedesoriano [Fedesoriano, 2021]. Although commonly applied to heart disease classification, we approach it here as a regression task.

The response variable is Age, and all other available variables are used as predictors. Our goal is to compare the performance of multiple linear regression (MLR) and ridge regression in predicting age, and to explore how regularization impacts model accuracy and complexity. We split the dataset into training and test sets to evaluate the models in a valid way.

The purpose of this project is to apply statistical learning methods covered in the course to a real dataset, and to assess which modeling approach is more suitable in this context. The analysis includes model building, assumption checking, performance evaluation, and a brief investigation of the most influential predictors. Our primary focus in this project is on prediction quality.

Descriptive data analysis/statistics

```
data <- read.csv("heart.csv")
```

To better understand the variables and their relationships, we perform descriptive data analysis on the heart failure dataset. Since our goal is to predict a person's age, we focus on identifying variables that might be relevant predictors.

Summary statistics

```r
# Mean, SD, Min, Max etc. across numeric variables

# Summary statistics for numeric variables
summary_stats <- data %>%
  summarise(across(where(is.numeric), list(
    Mean = ~mean(.),
    Median = ~median(.),
    SD   = ~sd(.),
    Variance = ~var(.),
    Min  = ~min(.),
    Max  = ~max(.)
  ), .names = "{.col}_{.fn}"))

# Table format
summary_stats_long <- summary_stats %>%
  pivot_longer(cols = everything(),
               names_to = c("Variable", "Statistic"),
               names_sep = "_") %>%
  pivot_wider(names_from = Statistic, values_from = value)

kable(summary_stats_long, caption = "Summary statistics for the numeric variables")
```

Table 1: Summary statistics for the numeric variables

Variable	Mean	Median	SD	Variance	Min	Max
Age	53.5108932	54.0	9.4326165	8.897425e+01	28.0	77.0
RestingBP	132.3965142	130.0	18.5141541	3.427739e+02	0.0	200.0
Cholesterol	198.7995643	223.0	109.3841446	1.196489e+04	0.0	603.0
FastingBS	0.2331155	0.0	0.4230456	1.789676e-01	0.0	1.0
MaxHR	136.8093682	138.0	25.4603341	6.482286e+02	60.0	202.0
Oldpeak	0.8873638	0.6	1.0665702	1.137572e+00	-2.6	6.2
HeartDisease	0.5533769	1.0	0.4974137	2.474204e-01	0.0	1.0

The table above presents key statistics for all numeric variables. We observe, for example, that Age ranges from 28 to 77, and MaxHR (maximum heart rate) has a relatively high standard deviation, indicating strong variability between individuals. These values give an initial sense of scale, spread, and potential outliers in the data, which are important to consider before modeling.

Age distribution

```r
ggplot(data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Distribution of Age", x = "Age", y = "Count")
```

Distribution of Age
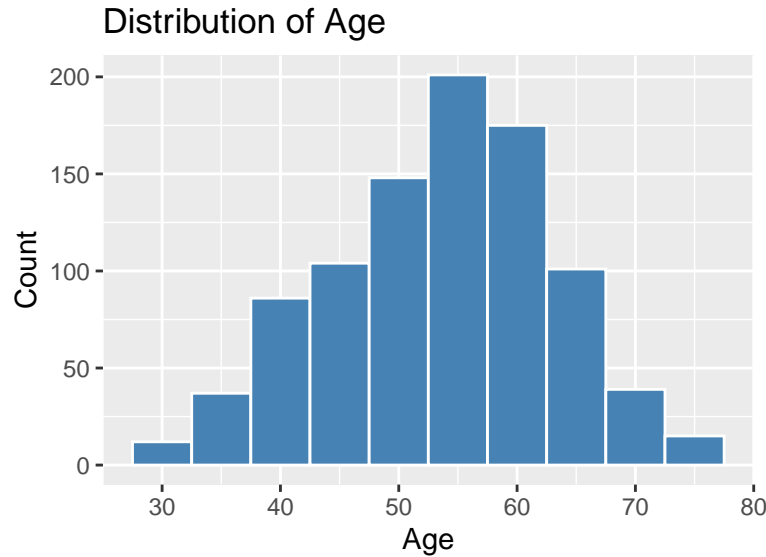
The distribution of Age is concentrated in the range 45–65, with a peak around 55 years. The distribution is roughly symmetric, with a few younger and older outliers. Since age is our target variable, this plot helps understand the range and whether any skew might influence model performance.

Correlation matrix

```
numeric_vars <- data[sapply(data, is.numeric)]

cor_matrix <- round(cor(numeric_vars), 4)

kable(cor_matrix, caption = "Correlation matrix of numeric variables")
```

Table 2: Correlation matrix of numeric variables

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1.0000	0.2544	-0.0953	0.1980	-0.3820	0.2586	0.2820
RestingBP	0.2544	1.0000	0.1009	0.0702	-0.1121	0.1648	0.1076
Cholesterol	-0.0953	0.1009	1.0000	-0.2610	0.2358	0.0501	-0.2327
FastingBS	0.1980	0.0702	-0.2610	1.0000	-0.1314	0.0527	0.2673
MaxHR	-0.3820	-0.1121	0.2358	-0.1314	1.0000	-0.1607	-0.4004
Oldpeak	0.2586	0.1648	0.0501	0.0527	-0.1607	1.0000	0.4040
HeartDisease	0.2820	0.1076	-0.2327	0.2673	-0.4004	0.4040	1.0000

The correlation matrix above highlights the relationships between numeric variables. We see a moderate negative correlation between Age and MaxHR (-0.38), and a mild positive correlation between Age and Oldpeak (0.26). Other variables, such as Cholesterol and RestingBP, appear weakly correlated with age. This suggests that MaxHR and Oldpeak could be useful predictors. The correlation between Age and HeartDisease is also not negligible, and it may be a useful predictor.

Relationship between numeric variables and age

Based on the correlation matrix, MaxHR and Oldpeak were the numeric variables most strongly related to age. While HeartDisease also appears as a numeric variable, it is a binary indicator and better treated as a categorical variable. Therefore, it was not included among the continuous variables explored through scatter plots, but rather in the analysis of the categorical variables.

```
# Numeric variables, scatter plots
ggplot(data, aes(x = Age, y = MaxHR)) + geom_point(alpha = 0.5) + geom_smooth(method = "lm", se = FALSE
  labs(title = "MaxHR vs Age", x = "Age", y = "Max Heart Rate")
```


MaxHR vs Age
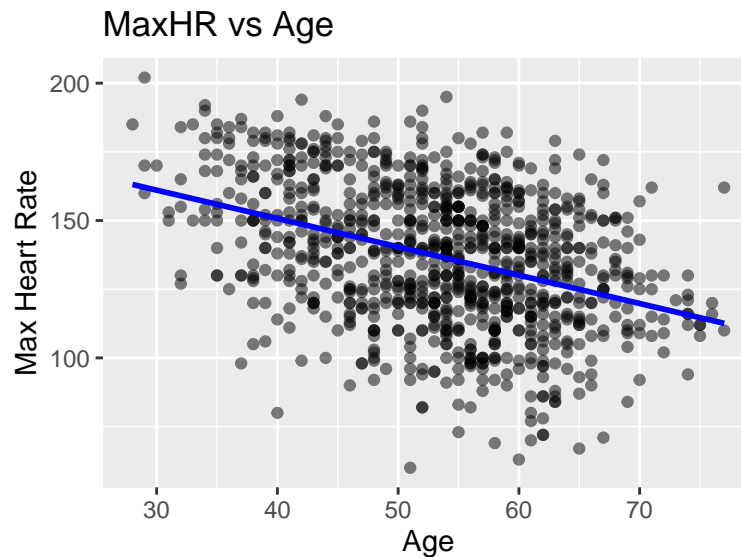
```
ggplot(data, aes(x = Age, y = Oldpeak)) + geom_point(alpha = 0.5) + geom_smooth(method = "lm", se = FALS
  labs(title = "Oldpeak vs Age", x = "Age", y = "Oldpeak")
```


Oldpeak vs Age
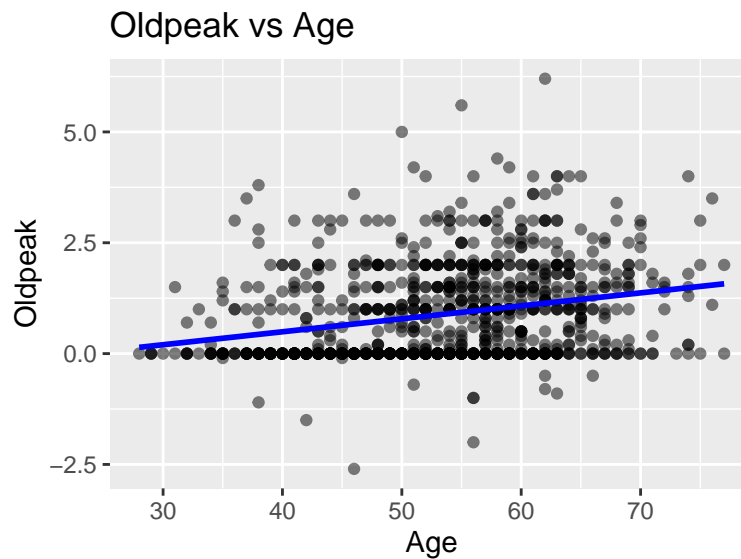
The scatter plot of MaxHR against Age reveals a clear negative linear trend: as age increases, maximum heart rate tends to decrease. This is physiologically expected and supports using MaxHR as a predictor. Oldpeak shows a slight upward trend with age, indicating a mild positive association. Both variables are therefore potentially useful in predicting age.

## Relationship between categorical variables and age

To assess the relationship between categorical variables and age, we plotted boxplots for each one. ExerciseAngina and ST_Slope showed the most distinct group differences in age and were therefore chosen as the most relevant to include. HeartDisease is also included because of the results found in the correlation matrix.

```
# Categorical variables, box plots
ggplot(data, aes(x = ExerciseAngina, y = Age)) + geom_boxplot() + labs(title = "Age by Exercise-Induced
```


Age by Exercise–Induced Angina
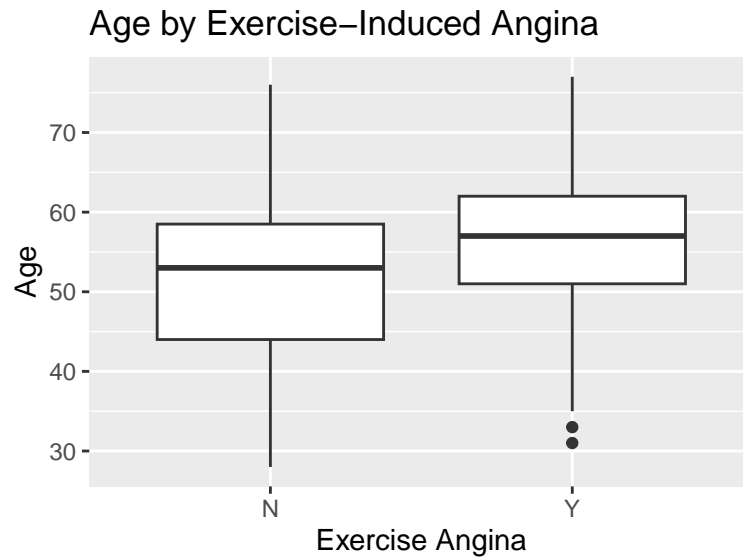
```
ggplot(data, aes(x = ST_Slope, y = Age)) + geom_boxplot() + labs(title = "Age by ST Slope", x = "ST Slop
```


Age by ST Slope
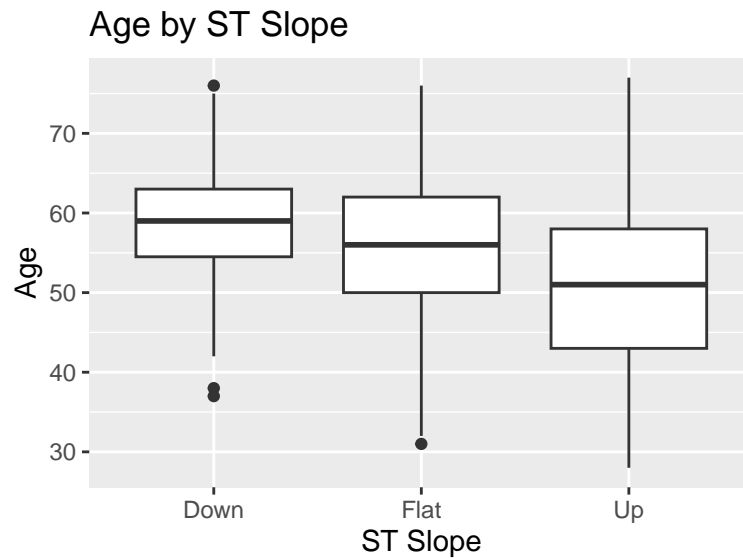
```
ggplot(data, aes(x = factor(HeartDisease), y = Age)) +
  geom_boxplot() +
  labs(title = "Age by Heart Disease Status", x = "Heart Disease (0 = No, 1 = Yes)", y = "Age")
```

Age by Heart Disease Status
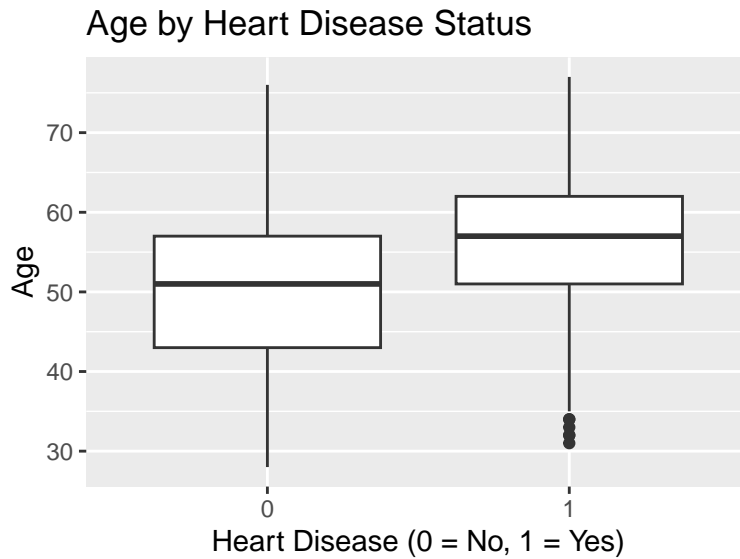
In the boxplot for ExerciseAngina, we observe that patients with angina induced by exercise (Y) tend to be older than those without (N). Similarly, the boxplot for ST_Slope shows that patients with a "Down" slope tend to be older than those with a "Flat" or "Up" slope. These differences suggest that both variables are potentially informative for modeling age and may capture patterns related to heart diseases. The boxplot for HeartDisease shows that patients with heart disease tend to be older on average than those without. This supports the idea that age is associated with heart disease presence.

Methods

We used the "Heart Failure Clinical Records" dataset and selected `Age` as the response variable. All other variables were used as predictors. Before we can apply the methods to the dataset, we need to split the data into training and test sets. Splitting the data ensures that we can train the model on one subset and evaluate its performance on unseen data, helping us assess how well the model generalizes. We randomly assign 70% of the data to the training set and the remaining 30% to the test set.

```
n <- nrow(data) # Number of observations

# Indexes for the training set (70% of the data)
train_idx <- sample(1:n, size = round(0.7 * n), replace = FALSE)

# Split the data
train_data <- data[train_idx, ]
test_data <- data[-train_idx, ]
```

Multiple Linear Regression

In this section, we will consider Multiple Linear Regression (MLR) as a method to predict age for the heart failure dataset. This choice of algorithm was due to the fact that MLR is easy to interpret. Another strength of using this method is its ability to extend to bivariate and categorical covariates. This suits the covariates we are evaluating, such as `Sex` and `ChestPainType`. With MLR we can also consider interaction and non-linear terms. However, a weakness of MLR is occurrence of collinearity when two or more predictors are correlated. This result in a reduced accuracy of the estimated regression.

Model Assumptions

First, we will consider the training data and assume that

$$\mathbf{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

where $\mathbf{Y}$ is a $(n \times 1)$-vector of responses (`Age`), X_j is the jth predictor, and β_j is the respective coefficient. From the training dataset, we have $n = 643$ patients, that is, 643 sampling units $(x_{1i}, \ldots, x_{pi}, y_i), 1 \le i \le n$, with $p = 15$ attributes for each patient. ϵ is a $(n \times 1)$-vector of random errors. We assume that $E(\epsilon_i) = 0$, that all ϵ_i have the same variance, and that ϵ_i are independent of each other and normally distributed.

To solve the prediction of age by using MLR, we used the `lm()`-function to estimate the regression coefficients β.

```
mlr_mod <- lm(Age ~ ., data = train_data)
```

We checked that the model assumptions of a MLR actually hold. To check this we used the model checking tools Tukey-Anscombe diagram and the QQ-diagram.

Evaluation of Multiple Linear Regression Performance

To get an idea of which the predictors $X_1, \ldots, X_n$ are useful in predicting the response, we checked if we could as well omit all predictor variables at the same time. We formulated the test as

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

vs.

$$H_1 : \text{at least one } \beta_j \text{ is non-zero.}$$

We used F-statistics to perform this test,

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)} \sim F_{p,(n-p-1)},$$

where TSS is the total sum of squares and RSS is the residual sum of squares. Under the assumptions we have made, F follows an $F_{p,(n-p-1)}$-distribution. The `lm()`-function calculates the F-statistics. If H_0 is true, we expect the F-statistics to be approximately equal to 1 [Muff, 2025].

To do the model selection between MLR and Ridge regression, we evaluate the model performance by Mean Squared Error (MSE) and R^2 (coefficient of determination). We use the test set to do this in a valid way. MSE measures the average of the squared differences between the actual and predicted values. It is calculated as:

$$\text{MSE} = \frac{1}{n} \sum_{i=n+1}^{n+q} (y_i - \hat{y}_i)^2$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and q is the number of patients in the test set. A low MSE indicates that the model's predictions are close to the actual values, making it a useful measure of model accuracy. R^2 measures the proportion of the variance in the dependent variable that is explained by the model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

The R^2 value ranges from 0 to 1, where higher values indicate better explanatory power.

```
# Model assessment
mlr_pred <- predict(mlr_mod, test_data)
mse_mlr <- mean((test_data$Age - mlr_pred)^2)

r2_mlr <- 1 - sum((test_data$Age - mlr_pred)^2) / sum((test_data$Age - mean(test_data$Age))^2)
```

Ridge regression

We now aim to predict **Age** using a different method, in order to compare the results with those obtained from multiple linear regression. We have chosen to use Ridge regression, both to gain a better understanding of this technique and to optimize the model through the use of a hyperparameter.

Ridge Regression is an extension of linear regression that includes a penalty for large coefficient values. The goal is to reduce model variance and improve generalization. Instead of minimizing only the Residual Sum of Squares (RSS), Ridge adds an additional term to the loss function:

$$\text{Loss} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Where:

- y_i are the observed values

- $\hat{y}_i$ are the model's predictions

- λ is a hyperparameter that controls the strength of the penalty on large coefficient values β_j

When $\lambda = 0$, the model is equivalent to ordinary linear regression.

In practice, Ridge regression shrinks the coefficients toward zero, but never exactly to zero— this means that all variables are retained in the model.

To implement the method to the Heart dataset, we used `model.matrix()` to create design matrices where all predictors are included. The categorical variables are automatically encoded as dummy variables. Ridge was implemented using the `glmnet` package in R, using $\alpha = 0$. Then, we used `cv.glmnet()` with 10-fold cross-validation to find the optimal value of lambda (the hyperparameter). The value `lambda.min` was used for the final models.

```
library(glmnet)

# Create design matrices
x_train <- model.matrix(Age ~ ., data = train_data)[, -1]
y_train <- train_data$Age
x_test <- model.matrix(Age ~ ., data = test_data)[, -1]
y_test <- test_data$Age

ridge_mod <- glmnet(x_train, y_train, alpha = 0) # `alpha=0` is the ridge penalty

# Cross-validation to find the best lambda
set.seed(123)
cv_ridge <- cv.glmnet(x_train, y_train, alpha = 0)
best_lambda <- cv_ridge$lambda.min
```

Ridge Regression - Advantages and disadvantages

Ridge regression is especially useful when two or more of the predictors are correlated with each other or when the number of predictors is large compared to the number of observations. The method helps prevent

8

overfitting by adding just enough bias to make the model's estimates more stable and robust. It also helps reduce model complexity without completely removing variables.

In our case, the most important advantages are that the method works well when the predictors are correlated. In our dataset, we can expect that variables such as `RestingBP`, `Cholesterol`, and `MaxHR` are related to each other. Additionally, Ridge regression is useful because we do not want to exclude any explanatory variables, but rather improve the precision of our predictions.

The general disadvantages of Ridge regression are that it includes all predictors in the final model, which can make interpretation challenging when the number of predictors is large. Although the penalty term shrinks all coefficients toward zero, it never sets any of them exactly to zero. This limitation can make Lasso a more suitable alternative in cases where variable selection and model simplification are important.

In our case, the number of predictors is not very large, and it is not necessarily beneficial to exclude any of them. Therefore, we argue that Ridge regression is well-suited for this particular dataset.

Evaluation of Ridge Regression Performance

To evaluate the performance of the model, we use Mean Squared Error (MSE) and R^2 (coefficient of determination) as performance metrics, in the same way as we did with MLR.

```r
# Predict Age on the test set
ridge_pred <- predict(cv_ridge, s = best_lambda, newx = x_test)

mse <- mean((y_test - ridge_pred)^2)
r2 <- 1 - sum((y_test - ridge_pred)^2) / sum((y_test - mean(y_test))^2)
```

Results and Interpretation

In this section, we will display the results of the methods on the training set and the evaluation of their performance using the test set. We will interpret the results and compare the models of which obtained the best results. In addition, we will discuss which variables had the highest impact on the prediction of age. The bias-variance trade off will also be taken into account.

Multiple Linear Regression

The MLR using the `lm()`-function is summarized in the following.

```
##
## Call:
## lm(formula = Age ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.9982  -5.5147   0.2984   5.4142  25.7556
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61.157004   3.597620  16.999  < 2e-16 ***
## SexM             -0.478152   0.832998  -0.574   0.5662
## ChestPainTypeATA -1.511517   0.998769  -1.513   0.1307
## ChestPainTypeNAP  1.150915   0.883941   1.302   0.1934
## ChestPainTypeTA   1.444529   1.589264   0.909   0.3637
## RestingBP         0.078874   0.017384   4.537 6.83e-06 ***
## Cholesterol      -0.004308   0.003231  -1.333   0.1829
## FastingBS         2.276642   0.808269   2.817   0.0050 **
```

```
## RestingECGNormal -4.375206   0.829786  -5.273 1.85e-07 ***
## RestingECGST     -2.149916   1.015703  -2.117   0.0347 *
## MaxHR            -0.110458   0.015006  -7.361 5.77e-13 ***
## ExerciseAnginaY   0.186851   0.807360   0.231   0.8171
## Oldpeak           1.142148   0.368024   3.103   0.0020 **
## ST_SlopeFlat     -1.258493   1.431131  -0.879   0.3795
## ST_SlopeUp       -1.063116   1.617031  -0.657   0.5111
## HeartDisease      0.786363   0.968457   0.812   0.4171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.918 on 627 degrees of freedom
## Multiple R-squared:  0.2866, Adjusted R-squared:  0.2696
## F-statistic:  16.8 on 15 and 627 DF,  p-value: < 2.2e-16
```

We observe that there are just some of the estimated regression coefficients with small p-values, indicating that these may be useful in predicting age. The intercept seems to be quite important in the prediction, and so does `RestingBP`, `RestingECGNormal`, and `MaxHR`.

Checking Model Assumptions

The model checking tools Tukey-Anscombe diagram and the QQ-diagram gave the following plots shown in the Figures 1 and 2.
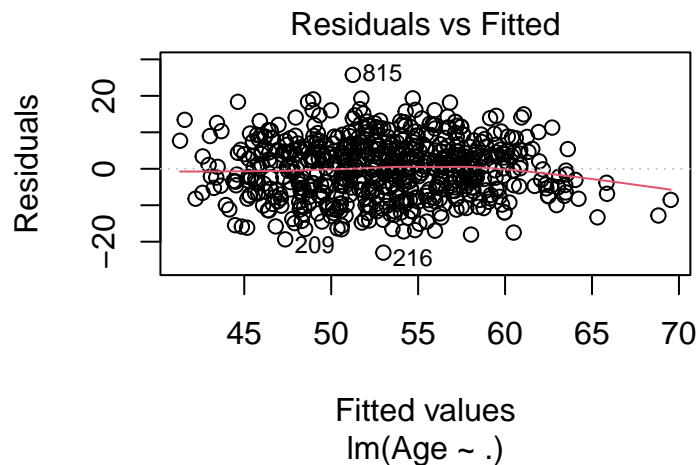

Figure 1: Tukey-Anscombe Diagram.

We observe in Figure 1 that the residuals seem to be dispersed along a horizontal line with no specific pattern. This imply that there is a linear relationship between the `Age` and the covariates, which we assume in MLR. We also notice that the residuals are spread around zero, which we assumed to be the mean value. In addition, it seems that the residuals have the same variance approximately being $\sqrt{20}$.

We observe in Figure 2 that the points approximately seem to follow a straight line, which indicates that the data are normally distributed. We can conclude with the that the model assumptions are fulfilled and that we may use MLR on this data set.

Evaluation of the Muliple Linear Regression

The F-test is shown in the summary of the MLR. We observe from the R output that the F is higher than 1 and that the respective p-value is very small, which means the predictors are useful in predicting the response. This also coincides with the p-values for the estimates of the β-coefficients.
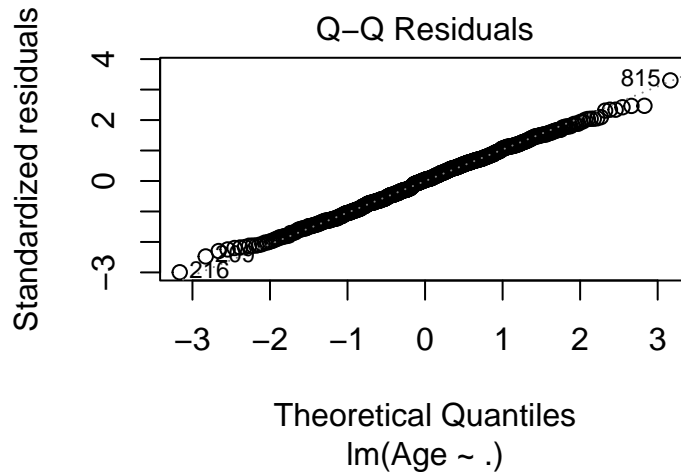
10

Figure 2: QQ-Diagram.

We have already noticed that not all p-values suggest significant coefficients. This means that perhaps only a subset of the predictors are useful. We could try to reformulate the test to check whether just some predictors are useful. In addition, we could try to extend the model to also consider interactions and non-linear terms. However, we would like to compare MLR to another model (Ridge regression), thus we will continue with the full model.

The MLR obtained a test MSE and a coefficient of determination approximately equal to 64.12 and 0.25, respectively. The model measures on average a difference of $\sqrt{64.12} \approx 8$ years between the estimated values of Age and the true value. In addition, 25% variability in the data is explained by the model.

Ridge Regression

To assess which variables have the greatest impact on the model's predictions, we can examine the magnitude of the coefficients in the Ridge regression model. Since Ridge does not set any coefficients exactly to zero, all predictors are included in the model, although some have a much stronger effect than others.
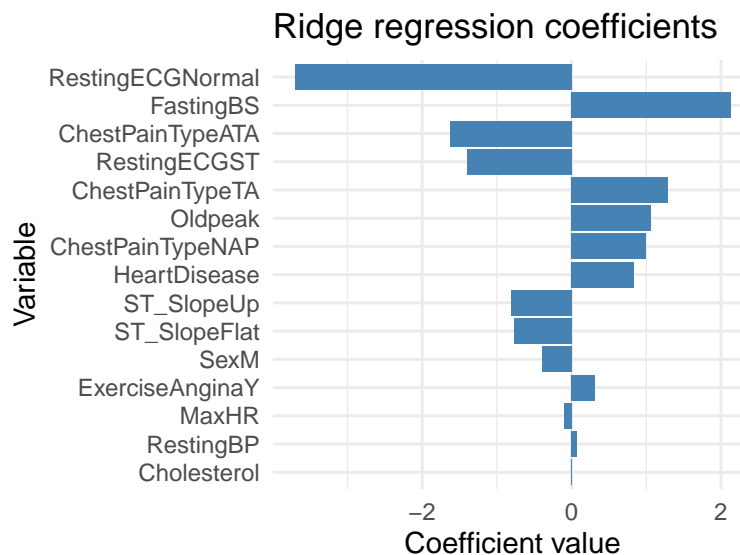

The variables with the largest absolute coefficients are those that influence age the most in the model, as you can see in the figure. RestingECGNormal (negative effect), FastingBS (positive effect) and ChestPainTypeNAP (positive effect) appeared to have the greatest influence. A positive effect means that as the value of the

variable increases, the predicted value of the response variable also increases. The opposite is true for a negative effect.

Optimal Value of the Shrinkage Parameter

Using 10-fold cross-validation with `cv.glmnet()`, the optimal value for the shrinkage parameter $\lambda = 1.195248$. This means that the model performed best, on average, when using this value for λ.
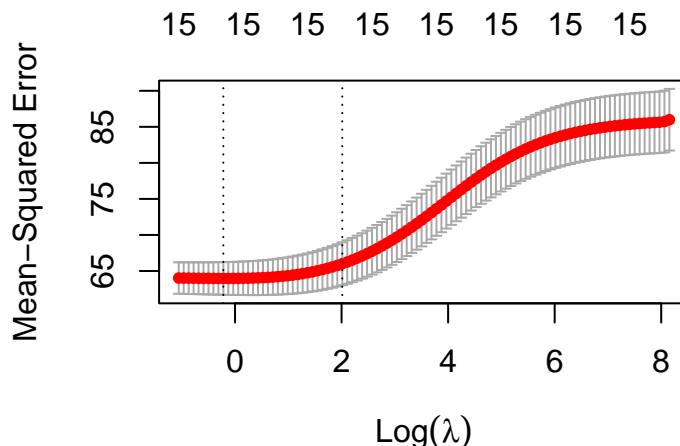
The figure shows the average Mean Squared Error (MSE) across different values of λ, with the horizontal axis representing $\log(\lambda)$. The red curve represents the average MSE for each value of λ, while the grey lines illustrate the standard error.

The optimal value of λ, i.e., the one that yields the lowest average cross-validated MSE, is in this case $\log(1.195248) \approx 0.18$, which indicates that the model performs best with a moderate amount of regularization.

Evaluation of the Ridge Regression

The Ridge regression model achieved a test MSE $= 70.92$ and a coefficient of determination $R^2 = 0.27$. This means that, on average, the model's predictions are off by about $\sqrt{70.92} \approx 8$ years, and it explains roughly 27% of the variation in age.

Nevertheless, the results show that Ridge regression is able to capture some underlying patterns and performs as expected by helping to control overfitting. The relatively low R^2 also highlights the importance of variable selection and data quality when building predictive models.

Model Selection

Both models achieved relatively similar results on the test set, with an average prediction error of approximately 8 years. Ridge regression resulted in a slightly higher MSE (70.92 vs. 64.12), but also a marginally better explanatory power measured by R^2 (0.27 vs. 0.25). This suggests that Ridge is able to capture slightly more of the variation in age, although the improvement is modest.

MLR can be sensitive when predictor variables are correlated — which can lead to high variance and unstable coefficient estimates. Ridge Regression addresses this by adding a penalty for large coefficients, which reduces variance and makes the model more robust to new data. This comes at the cost of a small increase in bias, but overall provides better generalization.

MLR is easy to interpret, particularly when it comes to understanding the effect of each individual predictor. Ridge Regression, on the other hand, is less interpretable, as all variables are retained in the model and the coefficients are jointly adjusted.

Since the dataset contains several predictors that are partially correlated, and we do not wish to exclude any variables, Ridge Regression is a reasonable choice. It provides better control over overfitting and handles

multicollinearity effectively. When the goal is accurate prediction rather than interpretation of individual coefficients, Ridge is preferred over MLR.

Summary

Both models provide a reasonable prediction of age, but neither MLR nor Ridge achieves high precision. While some variables are somewhat related to age, the dataset is not particularly well-suited for accurate age prediction. This is perhaps not surprising, as many of the variables (e.g., blood pressure, cholesterol, etc.) are influenced by both age and numerous other factors — and there is no variable in the dataset that directly indicates age. One way to improve the model would be to add predictors that are strongly correlated with age. Nevertheless, Ridge Regression offers slightly better generalization and is therefore the preferred choice in this context.

References

Fedesoriano. (2021). *Heart Failure Prediction Dataset.* Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Muff, S. (2025). Module 3: Linear Regression. *TMA4268 Statistical Learning.* Available at: https://github.com/stefaniemuff/statlearning2/blob/main/3LinReg/3LinReg.pdf (visited on 6th Apr. 2025)