Compulsory Exercise 2: Age Prediction for Heart Failure

Karianne Strand Bergem　Marte Ragnhild Hotvedt　Erlend Henriksen Winje

03 April, 2025

Abstract

This is the place for your abstract (max 350 words)

Introduction: Scope and purpose of your project

Problemstillingen vår kan være om vi kan predikere alder basert på de variablene i Heart Failure-datasettet? Er det vi skal finne ut av liksom

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

```
# 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")
```

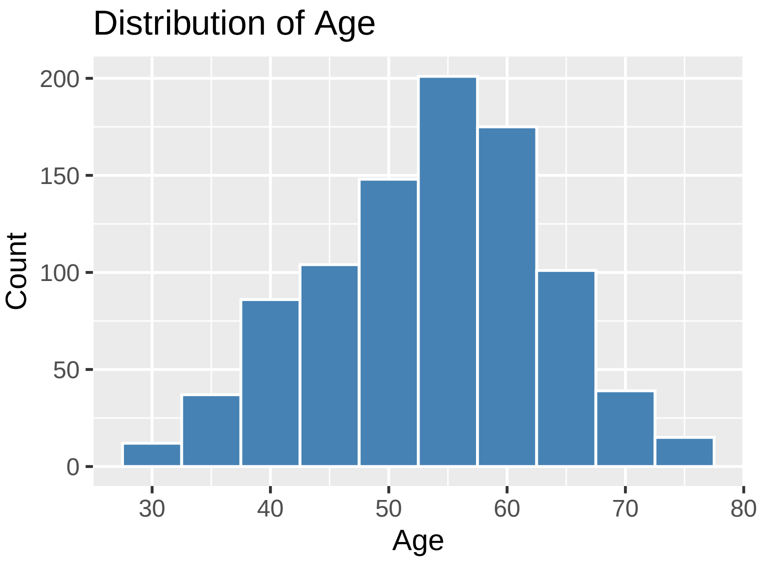
Table 1: Summary statistics for the numeric variables

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")
```



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

```r
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

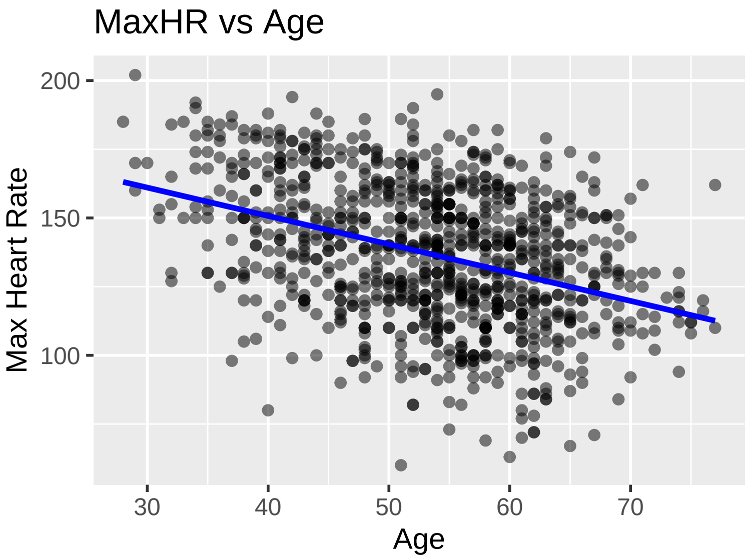
	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1.0000	0.2544	-0.0953	0.1980	-0.3820	0.2586	0.2820
RestingBP	0.2544	1.0000	0.1009	0.0702	-0.1121	0.1648	0.1076
Cholesterol	-0.0953	0.1009	1.0000	-0.2610	0.2358	0.0501	-0.2327
FastingBS	0.1980	0.0702	-0.2610	1.0000	-0.1314	0.0527	0.2673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

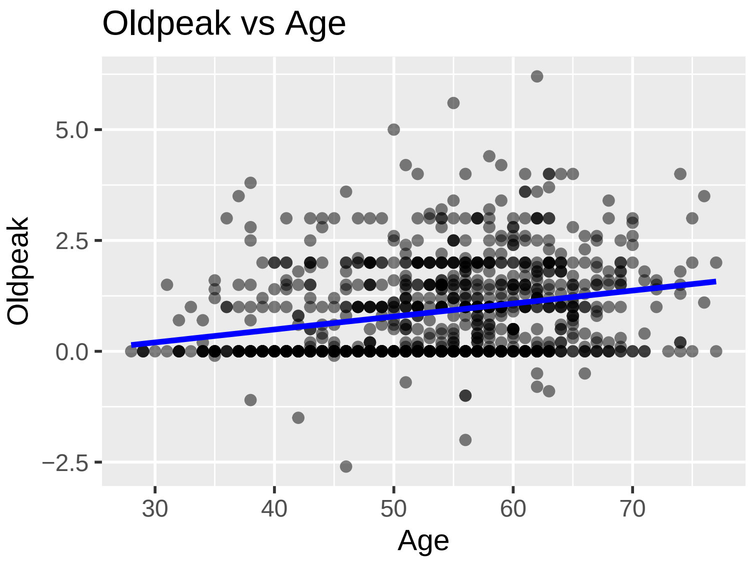
Relationship between numeric variables and age

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")
```



```
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")
```

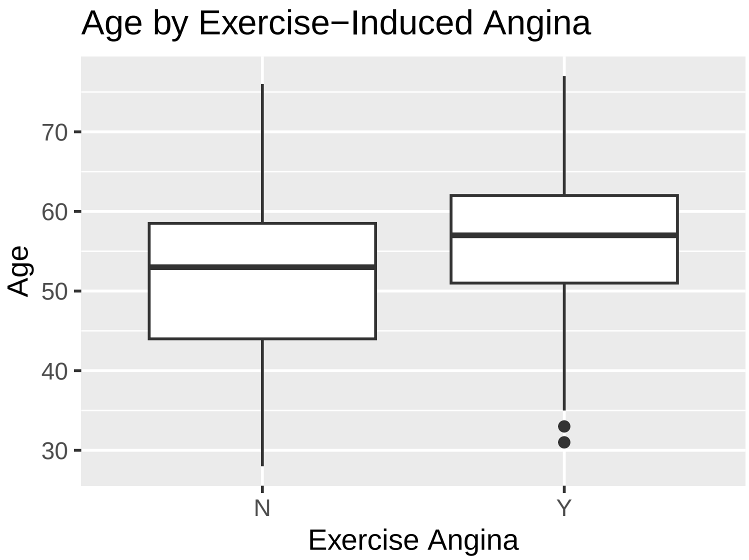


The scatter plot of MaxHR against Age reveals a clear negative linear trend: as age increases, maximum heart rate tends to decrease. This is physiologically expected and supports using MaxHR as a predictor. Oldpeak shows a slight upward trend with age, indicating a mild positive association. Both variables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

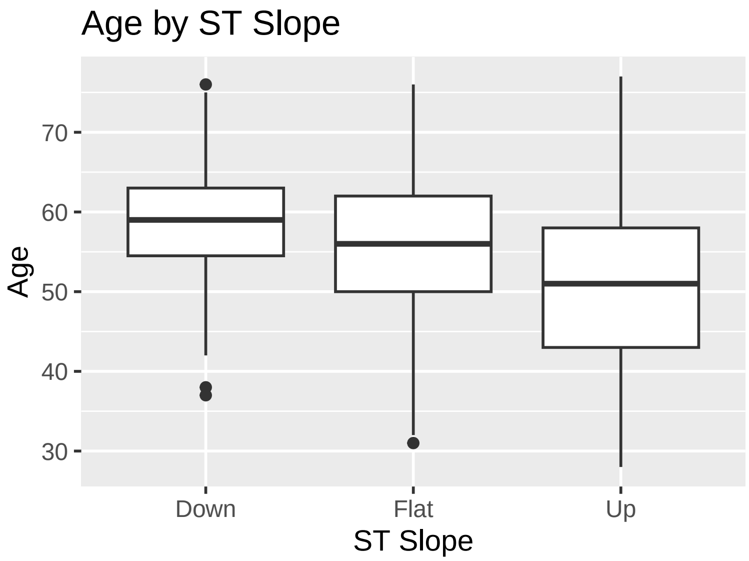
Relationship between categorical variables and age

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.

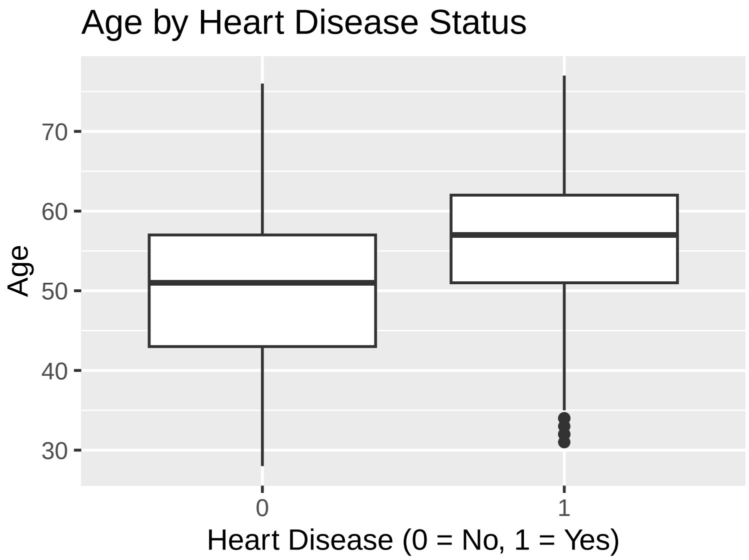
```
# Categorical variables, box plots
ggplot(data, aes(x = ExerciseAngina, y = Age)) + geom_boxplot() + labs(title = "Age by Exercise-Induced
```



```
ggplot(data, aes(x = ST_Slope, y = Age)) + geom_boxplot() + labs(title = "Age by ST Slope", x = "ST Slo
```



```r
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")
```



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.

```r
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 data set. This choice of method was due to the fact that MLR is easy to interpret. We assume that

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{Y}$ is a (918×1) vector of responses (age), β is a (16×1) of regression parameters, and ϵ is a (918×1) vector of random errors. $\mathbf{X}$ is the (918×16) design matrix given by

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,11} \\ 1 & x_{2,1} & \dots & x_{2,11} \\ \vdots & \dots & \dots & \vdots \\ 1 & x_{918,1} & \dots & x_{918,11} \end{bmatrix},$$

where row i contains a (16×1) vector x_i^T for observation i. The design matrix has full rank. In addition, since the number of rows is $n = 918$ and of columns is $(p + 1) = 16$, the model assumption $n >> (p + 1)$ holds. We assume that $\epsilon \sim N_{918}(0, \sigma^2 \mathbf{I})$.

The regression is done by the `lm()`-function:

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

```
## 
## 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
```

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

$$
\begin{gathered}
H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\
\text{vs.} \\
H_1 : \text{at least one } \beta_j \text{ is non-zero.}
\end{gathered}
$$

If H_0 is true, the F-statistics is expected to be ≈ 1. 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. However, we do notice that not all p-values suggest significant coefficients. This means that perhaps only a subset of the predictors are useful. To check this we can reformulate the test to

$$
\begin{gathered}
H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0 \\
\text{vs.} \\
H_1 : \text{at least one } \beta_j \text{ is non-zero,}
\end{gathered}
$$

where q is the number of coefficients β_j we want to test. We want to test if we can omit the coefficients with p-values higher than 0.05. To do this, we can perform the test by the `anova()`-function:

```r
mlr_mod_small <- lm(Age ~ ChestPainType + RestingBP + FastingBS + RestingECG
                    + MaxHR + Oldpeak + ST_Slope, data = train_data)
anova(mlr_mod_small, mlr_mod)
```

```
## Analysis of Variance Table
## 
## Model 1: Age ~ ChestPainType + RestingBP + FastingBS + RestingECG + MaxHR + 
##     Oldpeak + ST_Slope
## Model 2: Age ~ Sex + ChestPainType + RestingBP + Cholesterol + FastingBS + 
##     RestingECG + MaxHR + ExerciseAngina + Oldpeak + ST_Slope + 
##     HeartDisease
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    631 39492                           
## 2    627 39308  4    183.91 0.7334 0.5694
```

We observe a quite high p-value for the F-test, which means there is weak or even no evidence that the model gets better by including the coefficients with higher p-value than 0.05.

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

```
## [1] 70.6023
```

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

```
## [1] 0.2609674
```

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.

```r
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 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.

Mean Squared Error (MSE) measures the average of the squared differences between the actual and predicted values. It is calculated as:

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

where y_i is the actual value and $\hat{y}_i$ is the predicted value. 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.
In this project, we evaluate model performance on the test data using both MSE and R^2.

```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)
```

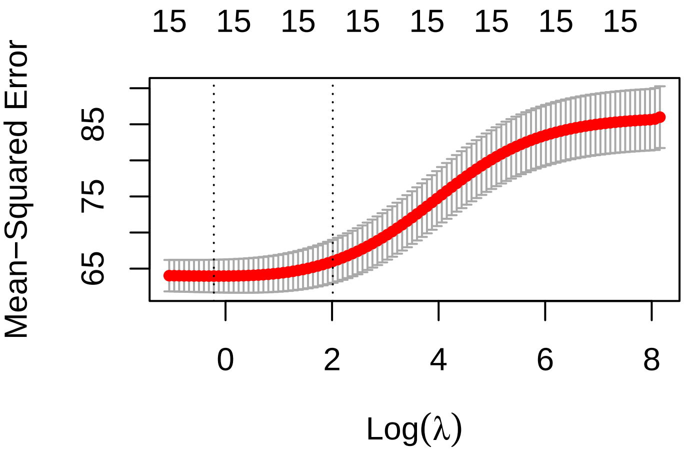
Results and interpretation

Evaluere modellene på testsettet. Sammenligne metodene – hvilken ga best resultater? Diskutere hvilke variabler som har størst betydning for prediksjon av alder.

Ridge Regression

Using 10-fold cross-validation with `cv.glmnet()`, the optimal value for the shrinkage parameter $\lambda = 1.195248$.

```r
plot(cv_ridge)
```



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.

This suggests that 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, 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.

Summary