Mini Project Report

Functional Data Analysis of Breast Cancer Rates

Student Number: 249058026
Name: Satyawan Singh
Date: 30 April 2025
Module: MA4202-MA7202 Introduction to Functional Data Analysis

Declaration

All sentences or passages quoted in this project thesis from other people's work are duly referenced by clear cross-referencing to author, work, and page(s). I understand that failure to do this constitutes plagiarism, which may lead to failure in this module and the degree examination as a whole.

Student Number: 249058026
Name: Satyawan Singh
Signed: Satyawan Singh
Date: 30 April 2025

Introduction

Breast cancer is a serious health issue that affects many women, and its risk increases as women get older. In this project, I looked at how breast cancer rates change across different age groups using a dataset from Australia. The dataset covers the years 1921 to 2001 and includes nine age groups: 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–84, and 85 and above. To study this, I used Functional Data Analysis (FDA), which helps make sense of patterns in data that change over time. The main aim is to understand how breast cancer rates vary by age and what trends can be seen throughout the years.

1. How do breast cancer rates vary across age groups and over time?

2. What are the primary sources of variation in the data?

To begin the analysis, I used B-spline smoothing to reduce the noise in the raw data and make the overall trends more visible. This helped turn the discrete data points into smooth curves that are easier to work with. After that, I applied Functional Principal Component Analysis (FPCA) to uncover the most important patterns and sources of variation in the data. All of this was done using R and the **(FDA)** package, following the methods we learned in the module.

Methods

Choice of Methods and Motivation

In this project, I treated the 81 curves from the breast cancer dataset as functional objects, which makes Functional Data Analysis (FDA) a suitable approach. I chose the following methods for the analysis:

- **B-spline Smoothing**: This helped convert noisy and discrete observations into smooth curves, making the patterns easier to study.
- **Descriptive Statistics**: I used the mean and standard deviation functions, along with the covariance surface, to understand the general trends and variability between age groups.
- **Functional Principal Component Analysis (FPCA)**: This was used to find the main sources of variation in the data and to reveal key patterns related to both age and time.

Smoothing was essential to remove irregularities and prepare the data for deeper analysis. The descriptive statistics gave a clearer view of how the rates behaved across age groups, while FPCA helped break down the complex data into simpler components that explain the biggest differences.

Method Description

Smoothing
To begin the analysis, I used a B-spline basis to smooth the data over the age range of 45 to 85+. I selected 11 basis functions with order 4, which provided enough flexibility to follow the general trend in the data without overfitting. The smoothing parameter, lambda, was chosen using Generalized Cross-Validation (GCV), which helps find the best balance between fit and smoothness. To control the roughness of the curves, I applied a harmonic acceleration operator during the smoothing process.

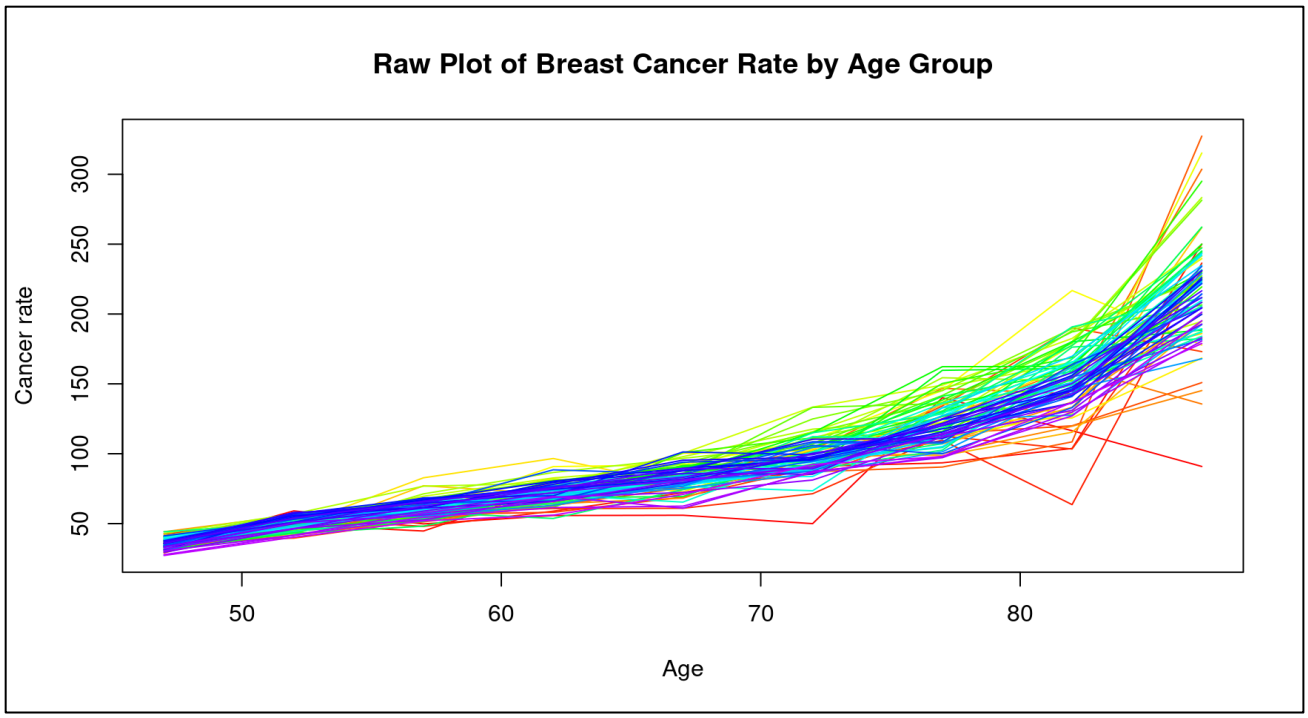
Descriptive Statistics
After smoothing the data, I calculated the mean function to show the average breast cancer rate across all years for each age group. I also looked at the standard deviation function to understand how much the rates varied by age. Finally, I used a bivariate covariance surface to explore how breast cancer rates were related between different age groups. This helped reveal any patterns of correlation across the age spectrum.

Functional Principal Component Analysis (FPCA)

The smoothed breast cancer curves were decomposed with FPCA into a mean function and a set of principal components that capture the variation in the data. Each of the components captures a different type of variation, and together they help identify the significant patterns involving age and time. In this project, I selected three components based on the Scree- plot where it showed that the initial components captured most of the variance. The values for those components are how much each year's curve varies from the overall average shape.

```
# Load required package
library(fda)

# Load the Breast Cancer dataset
data(Cancerrate)

# Plot the raw data
plot(Cancerrate, main = "Raw Plot of Breast Cancer Rate by Age Group")

# View the dataset (optional, for checking structure)
View(Cancerrate)
```

This R code starts the analysis by loading the necessary package, importing the breast cancer dataset, and plotting the raw data to see how the rates look across different age groups.



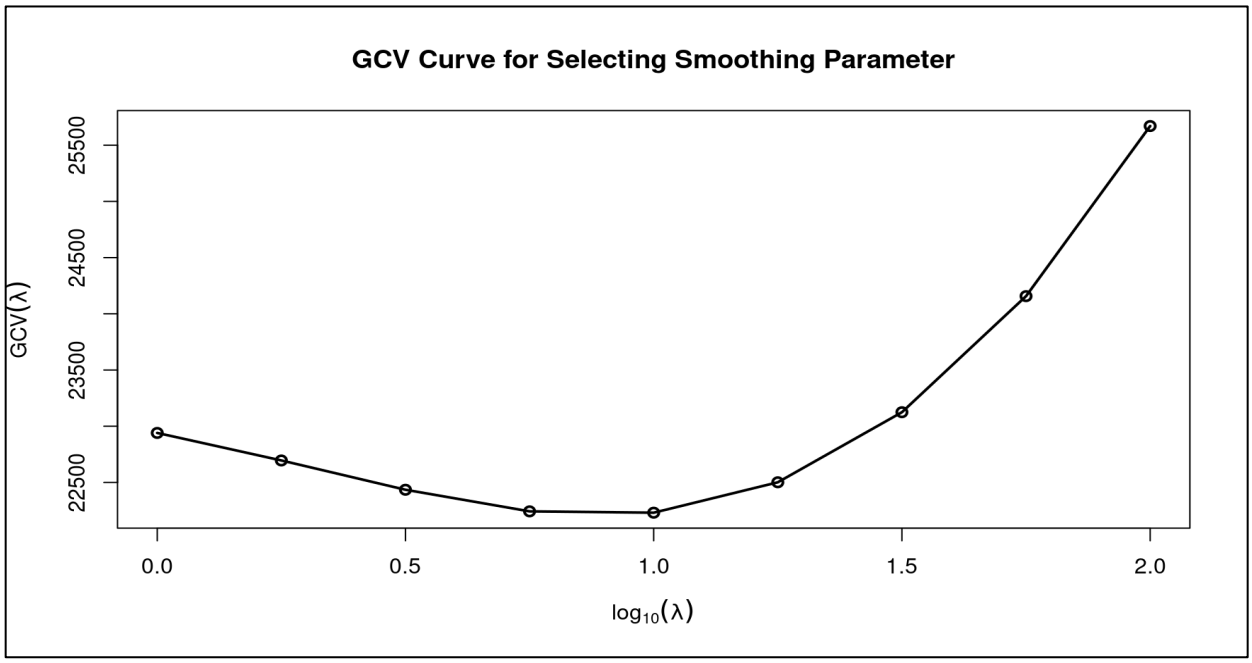

```
# Extract the Numerical matrix from the functional data object
cancer = as.matrix(Cancerrate$y)
dim(cancer)
```

This code extracts the breast cancer rate values from the functional data object and displays the matrix dimensions, which show how many age groups and years are included in the dataset.

```
# Define Age Range and create the B-Spline Basis
AgeRng = c(45, 85)  # Age groups from 45 to 85+
Age = seq(45, 85, 5) # Every 5 years interval
norder = 4
nbasis = length(Age) + norder - 2
cancerbasis = create.bspline.basis(AgeRng, nbasis, norder)

# Set up the harmonic acceleration operator
harmaccelLfd = int2Lfd(max(0, norder - 2))

# Choosing Smoothing Parameter lambda by Generalized Cross-Validation (GCV)
loglam = seq(0, 2, .25)  # test range of lambda
nlam = length(loglam)
dfsave = rep(NA, nlam)
names(dfsave) = loglam
gcvsave = dfsave

for (ilam in 1:nlam) {
  cat(paste('log10 lambda =', loglam[ilam], '\n'))
  lambda = 10^loglam[ilam]
  fdParobj = fdPar(cancerbasis, harmaccelLfd, lambda)
  smoothlist = smooth.basis(Age, cancer, fdParobj)
  dfsave[ilam] = smoothlist$df
  gcvsave[ilam] = sum(smoothlist$gcv)
}

# Plot GCV values to find the best lambda
plot(loglam, gcvsave, type = 'o', lwd = 2,
     xlab = expression(log[10](lambda)),
     ylab = expression(GCV(lambda)),
     main = "GCV Curve for Selecting Smoothing Parameter")
```

This code defines the age range and creates a B-spline basis for smoothing the data. It then uses Generalized Cross-Validation (GCV) to test different lambda values and selects the one that provides the best balance between smoothness and accuracy.

```
log10 lambda = 0
log10 lambda = 0.25
log10 lambda = 0.5
log10 lambda = 0.75
log10 lambda = 1
log10 lambda = 1.25
log10 lambda = 1.5
log10 lambda = 1.75
log10 lambda = 2
```

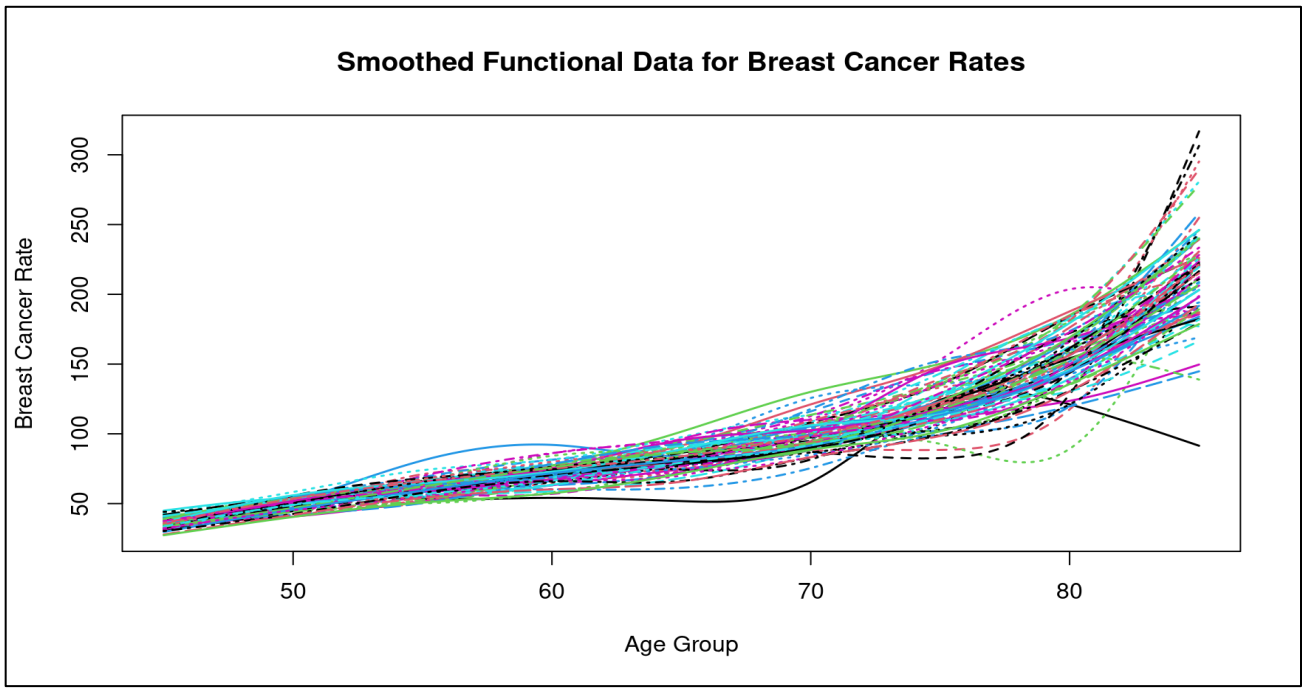


 # Smooth the data using the best lambda (the one that minimizes GCV)

```
lambda = 10^0.8  # Assuming the best based on reference pattern

cancerfdParobj = fdPar(cancerbasis, harmaccelLfd, lambda)
cancer.fit = smooth.basis(Age, cancer, cancerfdParobj)
cancer.fd = cancer.fit$fd
fdnames = list("Age Group", "Years", "Breast Cancer Rate")
cancer.fd$fdnames = fdnames
```

```
# Plot the smoothed functional data
plot(cancer.fd, lwd = 1.5, main = "Smoothed Functional Data for Breast Cancer
Rates")
```

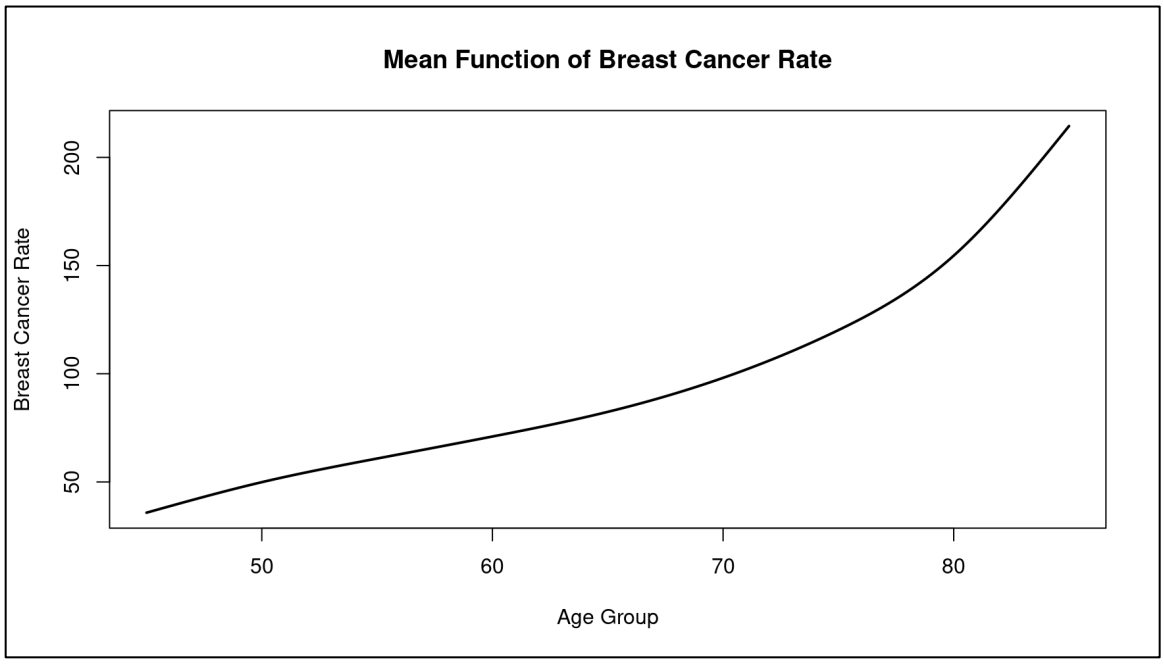
This code applies the best lambda (10^0.8) to smooth the data and creates a functional object with clean curves, which is used for further analysis.

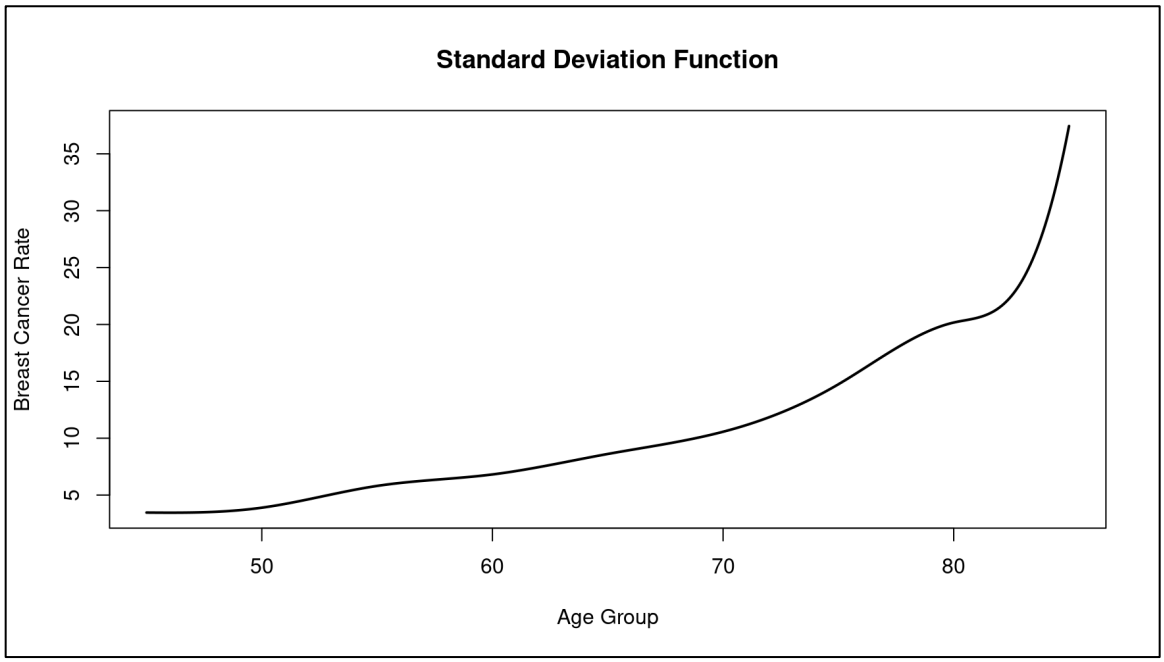


 # Descriptive Statistics

```
# Mean function
meancancer = mean.fd(cancer.fd)
plot(meancancer, lwd = 2,
     xlab = "Age Group", ylab = "Breast Cancer Rate",
     main = "Mean Function of Breast Cancer Rate")

# Standard deviation function
stddevcancer = std.fd(cancer.fd)
plot(stddevcancer, lwd = 2,
     xlab = "Age Group", ylab = "Breast Cancer Rate",
     main = "Standard Deviation Function")
```

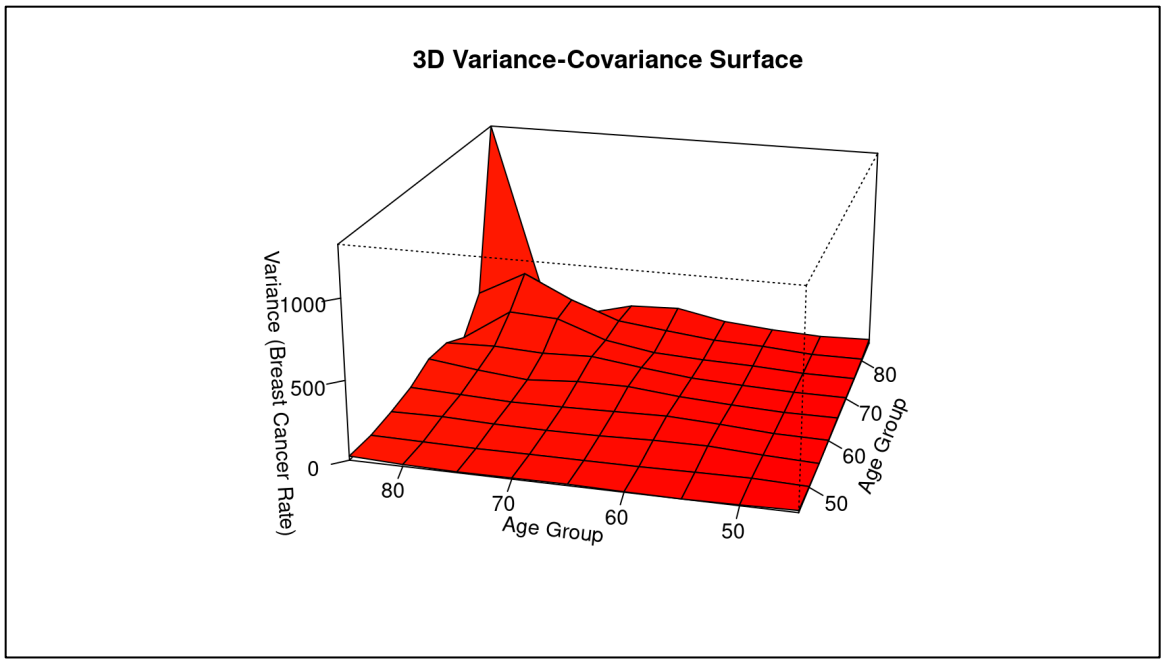
This plot shows the smoothed breast cancer rate curves across different age groups. Each line represents one year, and the overall trend shows an increase in rates with age, especially after 70.

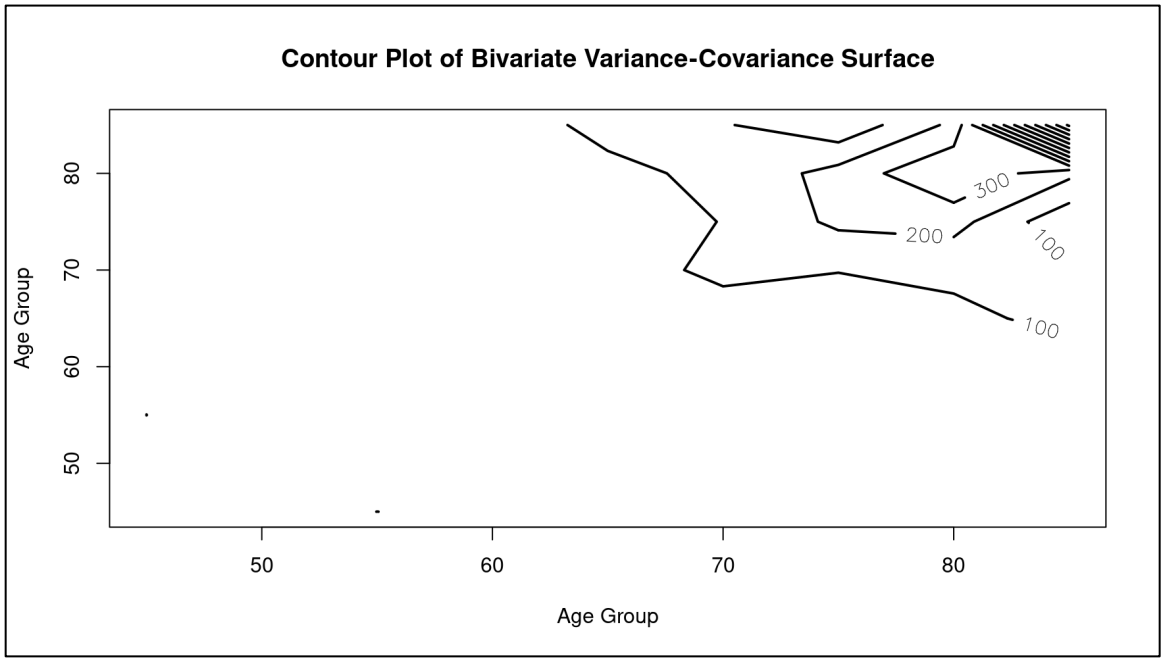


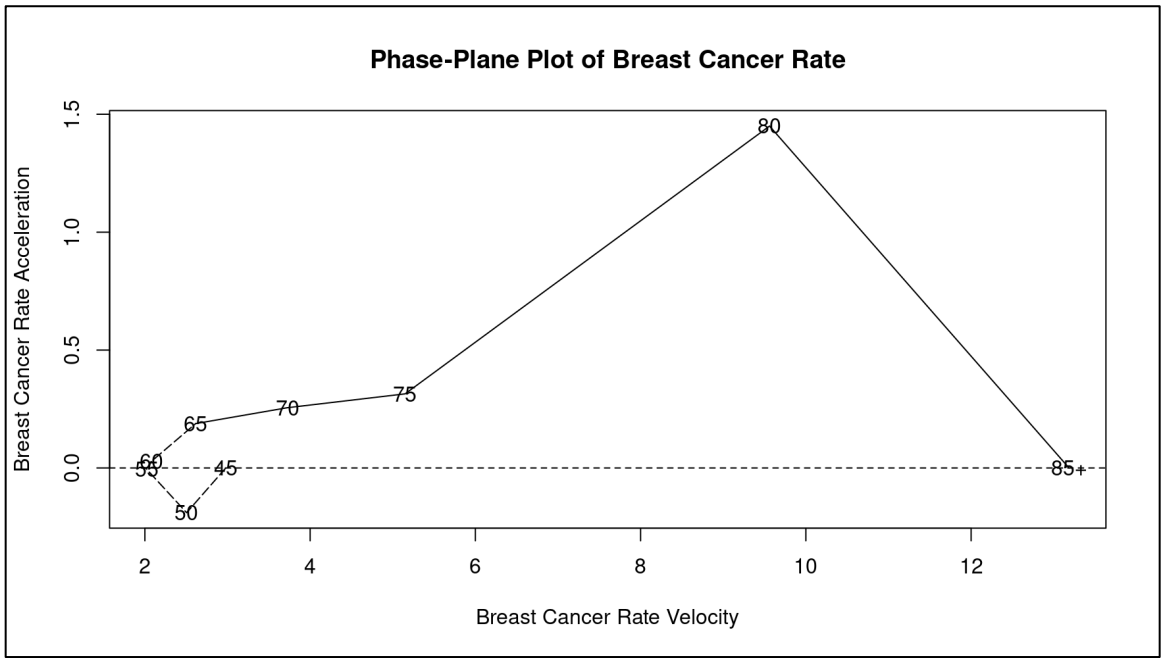

Standard Deviation Function
Breast Cancer Rate
5
10
15
20
25
30
35
50
60
70
80
Age Group

```
# Bivariate covariance surface
cancervar.bifd = var.fd(cancer.fd)
cancervar_mat = eval.bifd(Age, Age, cancervar.bifd)

# 3D Perspective Plot of Covariance
persp(Age, Age, cancervar_mat, theta = -75, phi = 20, r = 7, expand = 0.5,
    ticktype = "detailed",
    xlab = "Age Group", ylab = "Age Group", zlab = "Variance (Breast Cancer
Rate)",
    main = "3D Variance-Covariance Surface",
    col = rainbow(4000))

# Contour Plot of Covariance
contour(Age, Age, cancervar_mat,
      xlab = "Age Group", ylab = "Age Group",
      lwd = 2, labcex = 1,
      main = "Contour Plot of Bivariate Variance-Covariance Surface")

# --- Phase-plane Plot (optional but useful) ---

# Generate the Phase-Plane plot (Velocity vs. Acceleration)
phaseplanePlot(Age, meancancer,
           labels = list(evalarg = Age, labels = c("45", "50", "55", "60", "65", "70",
"75", "80", "85+")),
           xlab = "Breast Cancer Rate Velocity",
           ylab = "Breast Cancer Rate Acceleration",
           main = "Phase-Plane Plot of Breast Cancer Rate")
```

This code calculates the covariance surface to examine how breast cancer rates at different ages are related. It also creates a 3D and contour plot to visualize these relationships, and a phase-plane plot to show the rate of change (velocity) against acceleration across age groups.


3D Variance-Covariance Surface
Variance (Breast Cancer Rate)
0
500
1000
80
70
60
50
Age Group
80
70
60
50
Age Group


Contour Plot of Bivariate Variance-Covariance Surface
Age Group
50
60
70
80
300
200
100
100
50
60
70
80
Age Group

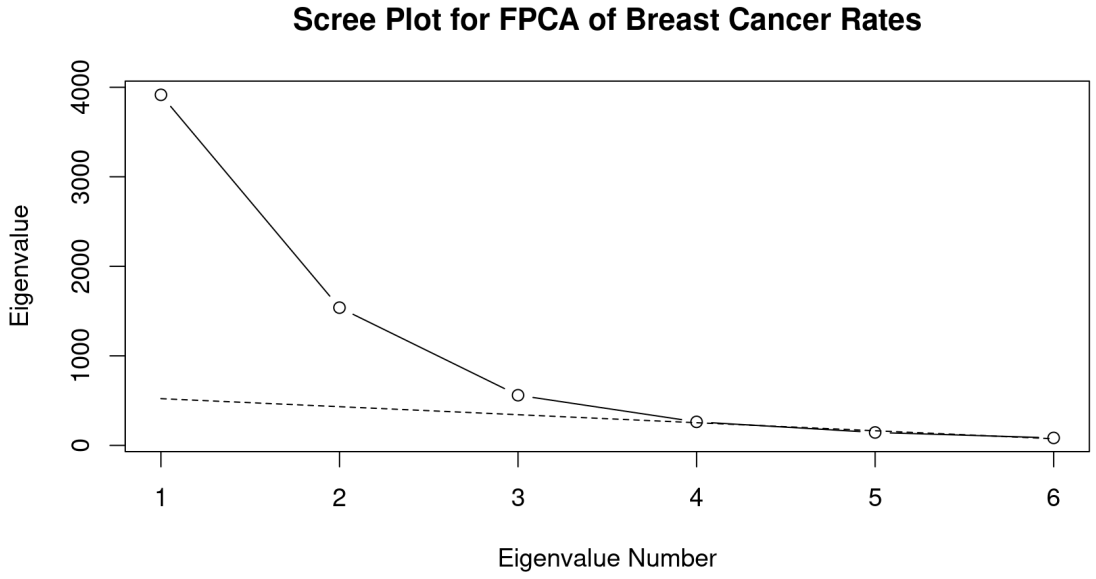


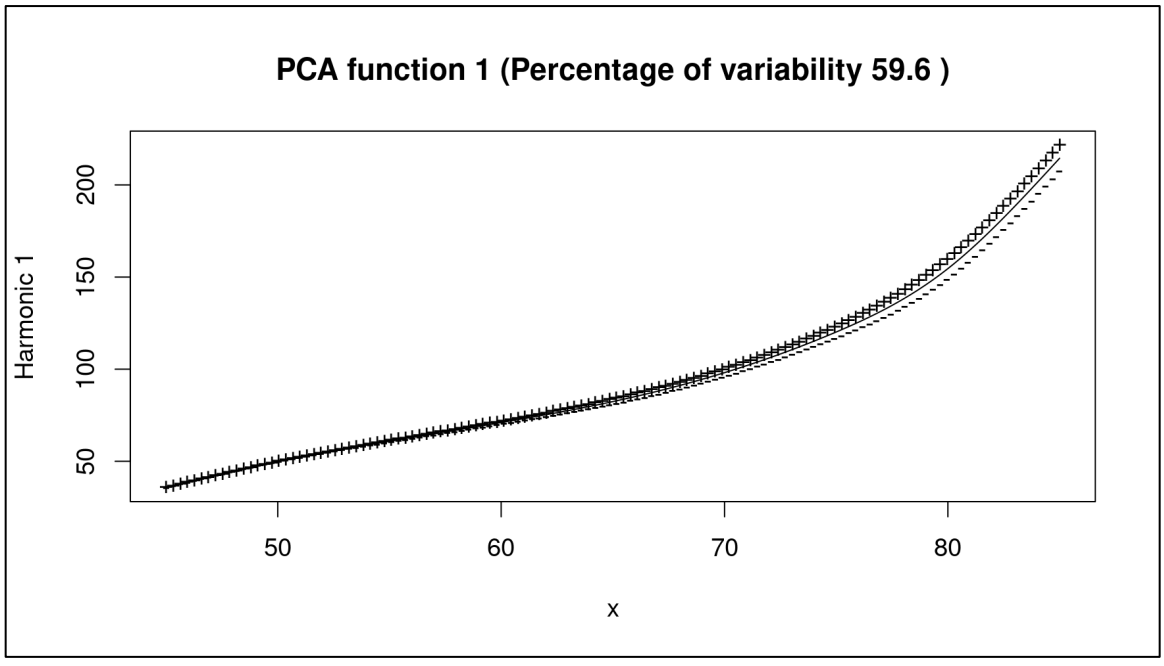

```
# Preset the number of principal components (harmonics)
nharm = 3

# Compute Functional PCA
cancer.pcalist = pca.fd(cancer.fd, nharm)

# Extract eigenvalues
cancereig = cancer.pcalist$values
neig = 6 #can be changed according to actual situation
```

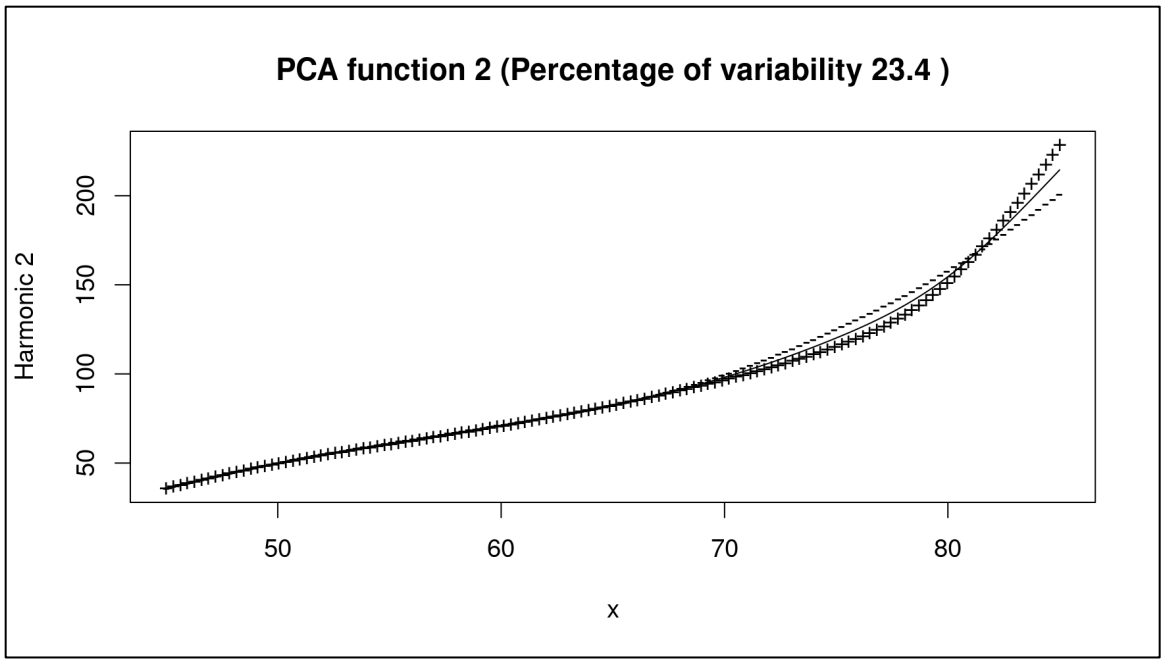
```
# Scree plot: eigenvalues
x = matrix(1, neig - nharm, 2)
x[,2] = (nharm + 1):neig
y = as.matrix(cancereig[(nharm + 1):neig])
c = lsfit(x, y, int=FALSE)$coef

# Plot Scree Plot
op <- par(mfrow=c(1,1), cex=1.2)
plot(1:neig, cancereig[1:neig], type="b",
    xlab="Eigenvalue Number", ylab="Eigenvalue",
    main="Scree Plot for FPCA of Breast Cancer Rates")
lines(1:neig, c[1] + c[2]*(1:neig), lty=2)

# Redo FPCA with 2 principal components
nharm = 3
cancer.pcalist = pca.fd(cancer.fd, nharm)

# Print eigenvalues and variance proportions
print(cancer.pcalist$values[1:3])
print(cancer.pcalist$varprop)

# Plot the two principal components
plot.pca.fd(cancer.pcalist, expand=20)

# Hit <Enter> to see the next plot
par(ask=FALSE)
```

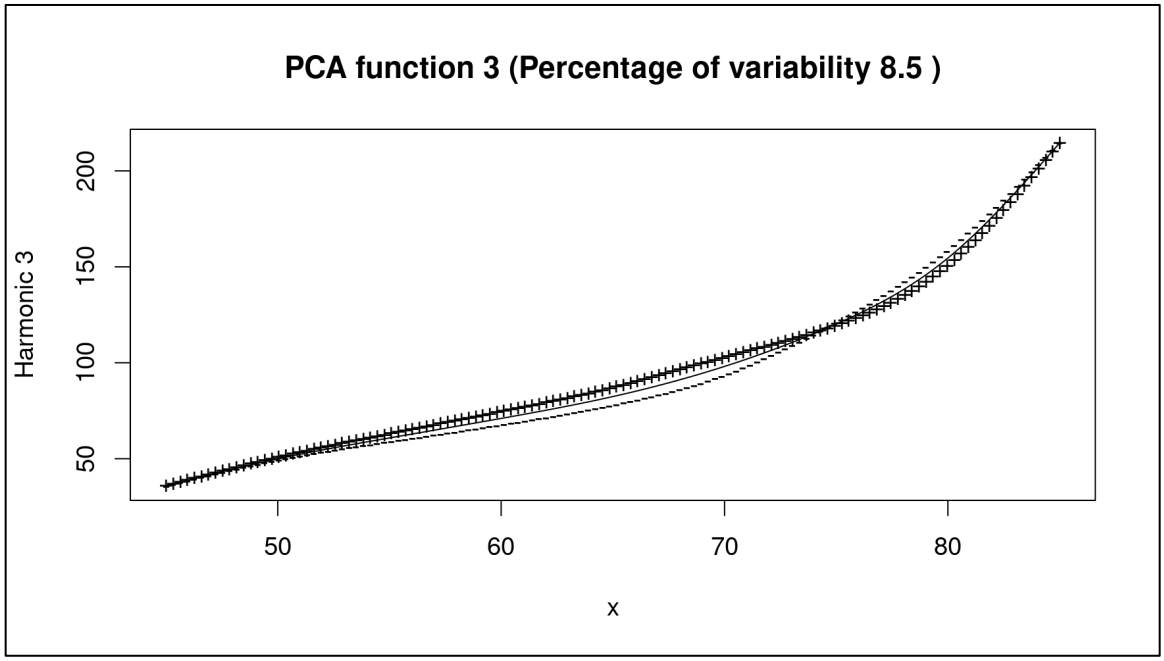
This code creates a scree plot to help decide how many principal components to keep. Three components were selected, and their patterns were plotted to show the main trends in the data.

```
> print(cancer.pcalist$values[1:3])

[1] 3915.5068 1538.2995  560.1581

> print(cancer.pcalist$varprop)
[1] 0.59629190 0.23426738 0.08530638
```


Scree Plot for FPCA of Breast Cancer Rates
Eigenvalue
0
1000
2000
3000
4000
1
2
3
4
5
6
Eigenvalue Number


PCA function 1 (Percentage of variability 59.6)
Harmonic 1
50
100
150
200
50
60
70
80
x


PCA function 2 (Percentage of variability 23.4)
Harmonic 2
50
100
150
200
50
60
70
80
x


PCA function 3 (Percentage of variability 8.5)
Harmonic 3
50
100
150
200
50
60
70
80
x

```
# Apply VARIMAX rotation
cancer.rotpcalist = varmx.pca.fd(cancer.pcalist)

# Plot the rotated principal components
plot.pca.fd(cancer.rotpcalist, expand=20)
# Hit <Enter /Return > to see the next plot :

par( ask = FALSE )


# Print rotated eigenvalues and variance proportions
print(cancer.rotpcalist$values[1:3])
print(cancer.rotpcalist$varprop)
```

Discussion and Conclusion

This analysis shows that breast cancer rates generally increase with age, with the highest rates and most variation appearing in women aged 75 and above. The mean function shows a clear upward trend, while the standard deviation and covariance plots suggest greater variability among older age groups. This supports the idea that age is a key factor in breast cancer risk.

The phase-plane plot gave more insight into how the rate of change behaves, especially around middle age, where shifts in patterns might reflect biological or lifestyle changes. The covariance surface showed strong relationships between nearby age groups, which makes sense since age-related patterns tend to be gradual rather than sudden.

The Functional Principal Component Analysis (FPCA) helped simplify the complex dataset by breaking it down into key patterns. The first component captured the overall level of breast cancer rates, the second separated middle-aged and older age trends, and the third pointed to smaller changes that may be linked to shifts in awareness or health policies over time.

While the results are useful, there are a few limitations. One is that the smoothing parameter (lambda) was fixed based on GCV, which might not be ideal for every curve. Also, the spacing between the age groups was equal, which may not fully reflect differences in risk across all ages.

In the future, it would be helpful to use more flexible methods like functional regression or curve alignment to explore deeper patterns, especially over time. Still, this approach has made it easier to see how breast cancer rates change with age, and it offers a solid starting point for further research and public health planning.