Mathematics for Big Data - Exercise 2b

Lawrence Adu-Gyamfi (1484610)

10/06/2019

Contents

1 INTRODUCTION

This report presents the code and the results of functional data analysis of the medfly data.

The medfly data have been a popular dataset for functional data analysis.The data consist of records of the number of eggs laid by 50 fruit flies on each of 31 days, along with each individual???s total lifespan.

2 Question 1

The following section uses code provided in the assignment for smoothing the data for the number of eggs and choosing the smoothing parameter by GCV (Generalized Cross-Validation).

```
## The best lambda used for smoothing is:  54.59815
```

2.1 Question

Plot the smoothed data.

2.2 Solution

Below is a plot of the smoothed data showing the number of eggs laid by each of the different fruit flies as a function of time (day).

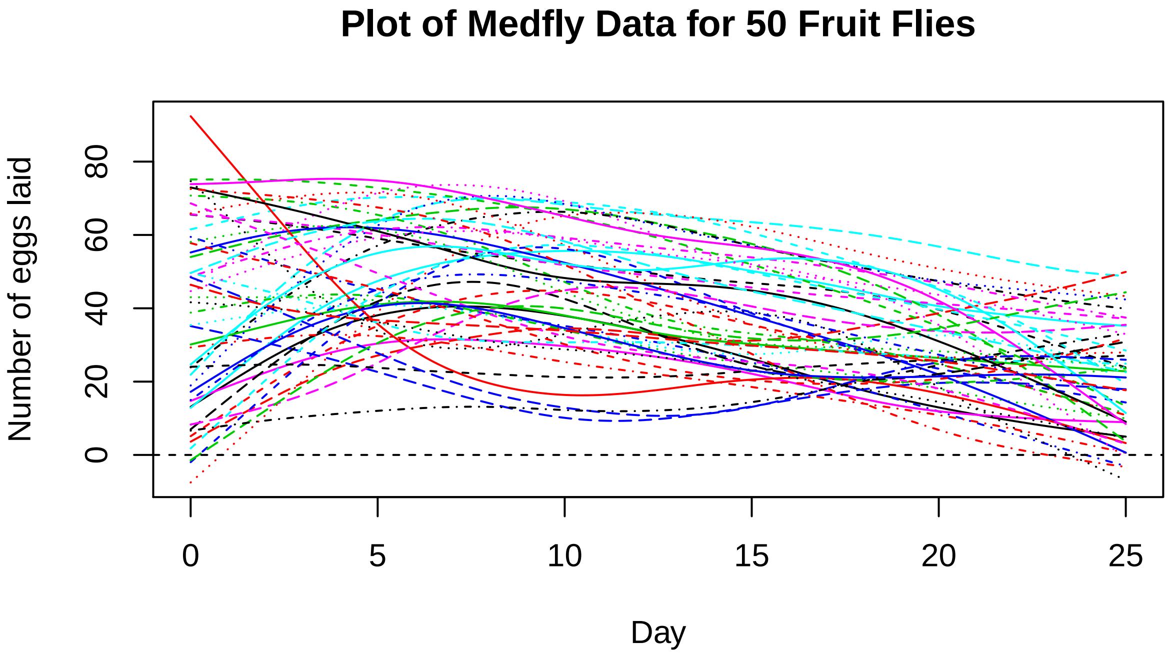


Figure 1: Plot of smoothed medfly dataset.

3 Question 2

3.1 Question

Conduct a principal components analysis using these smooths. Are the components inter- pretable? How many do you need to retain to recover 90% of the variation.

3.2 Solution

Below are the results of the explained variability for each of the principal components. The cummulated sum is shown as well to show many principal components help to recover most of the variability.

Table 1: PCA results

PC	Variability_Explained	CUMSUM
1	0.608	0.608
2	0.292	0.900
3	0.065	0.965
4	0.030	0.994
5	0.005	0.999
6	0.001	1.000
7	0.000	1.000
8	0.000	1.000
9	0.000	1.000
10	0.000	1.000

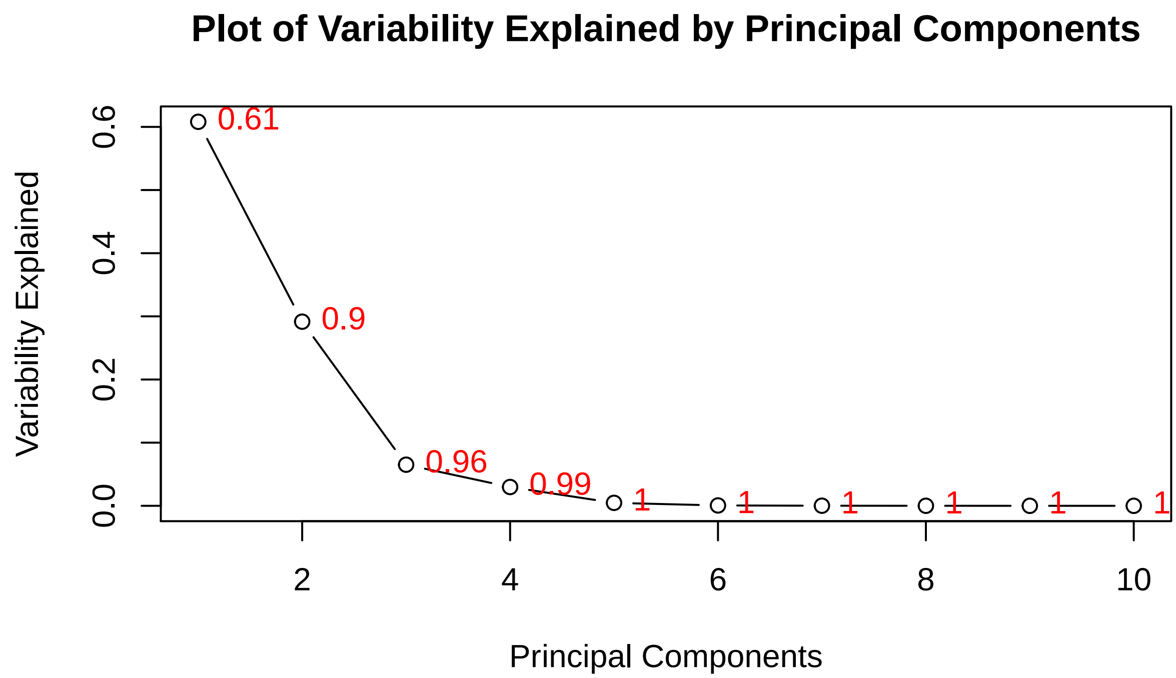


Figure 2: Plot of Variability Explained by Principal Components of Medfly Dataset

We are able to recover more than 90% of the variability using between 2 and 3 principal components, and by the 5th principal component, almost all the variability is recovered.

The following plots show the first 3 principal components.

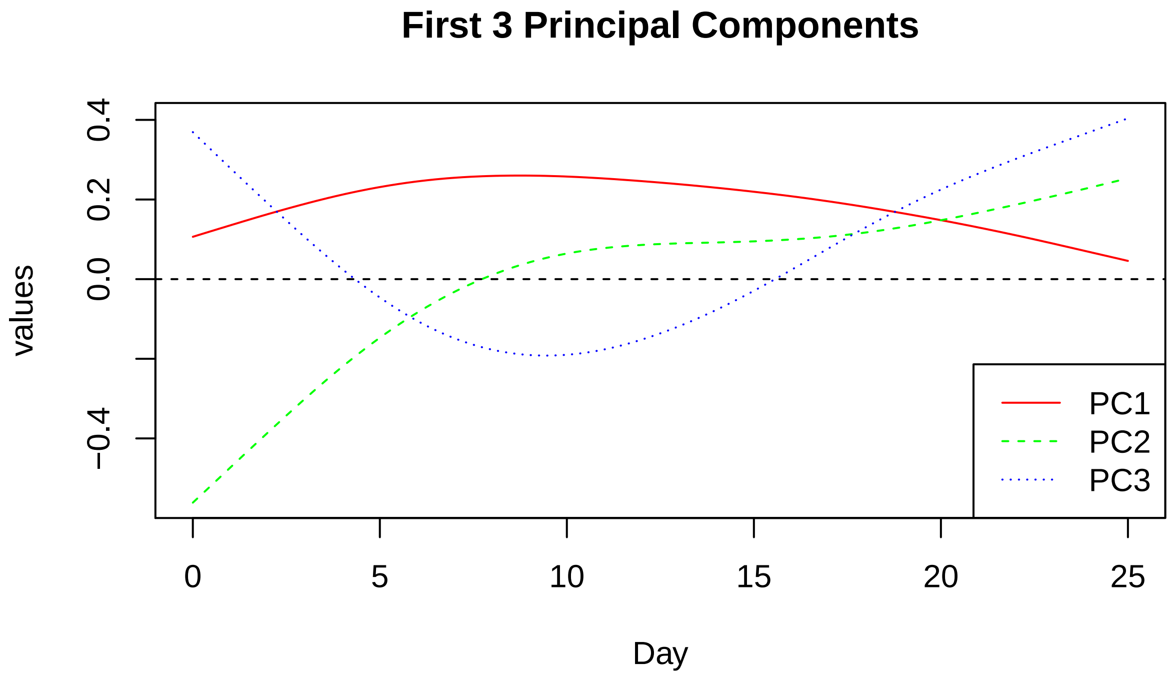


Figure 3: First 3 Principal Components

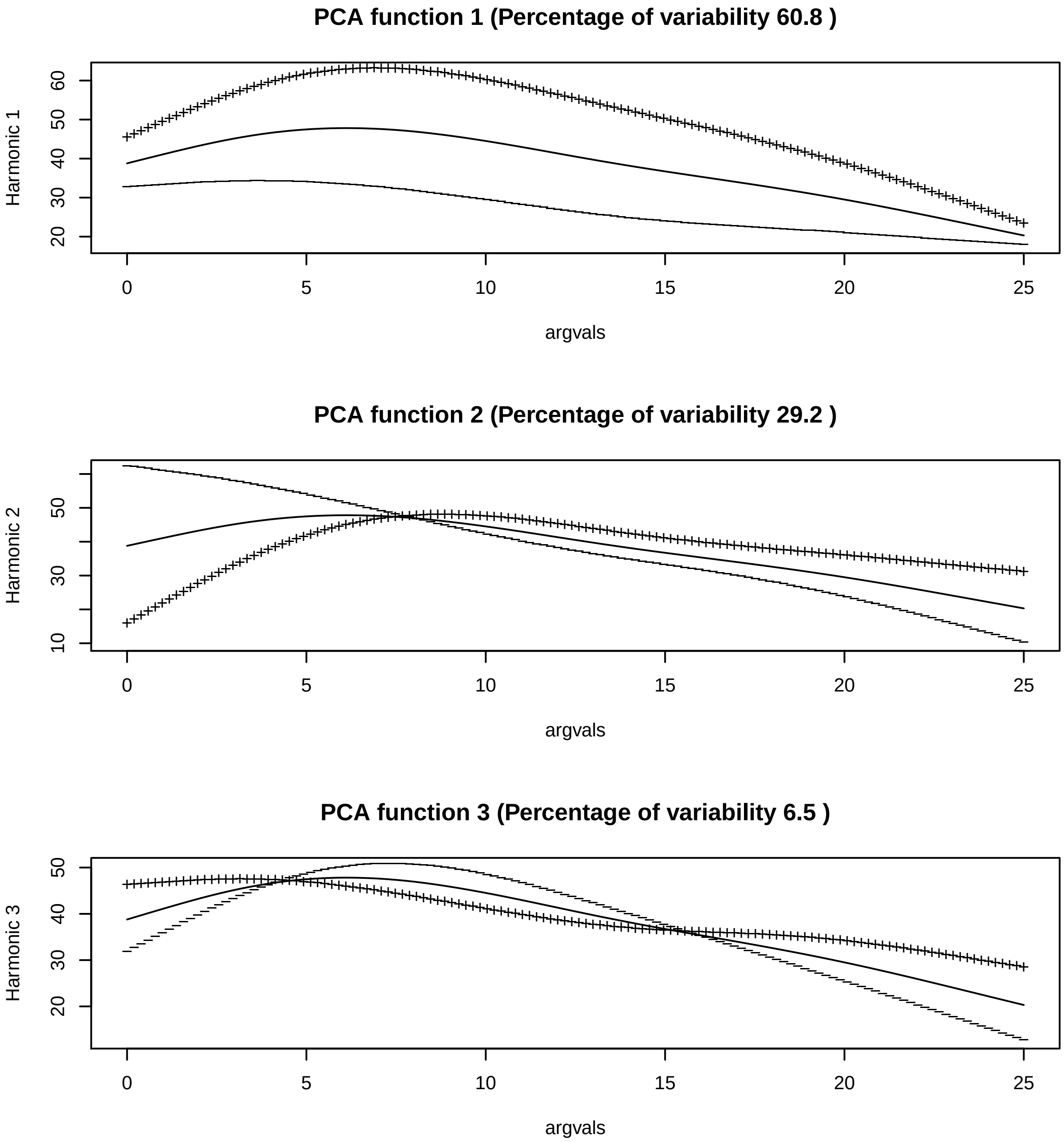


Figure 4: Individual Plots of first 3 Principal Components

Comparing the plots of the principal components to the plot of the entire dataset, the first principal component seems to describe the total eggs laid over the entire period for each day, which makes sense for this recover much of the variability of the dataset.

4 Question 3

4.1 Question

Perform a functional linear regression to predict the total lifespan of the fly from their egg laying. Choose a smoothing parameter by cross validation, and plot the coefficient function along with confidence intervals.

4.2 Solution

```
## Best lambda used for the regression is:  54.59815
```

The following plot shows the predicted values of total lifespan returned by the model against the actual values of the dataset.

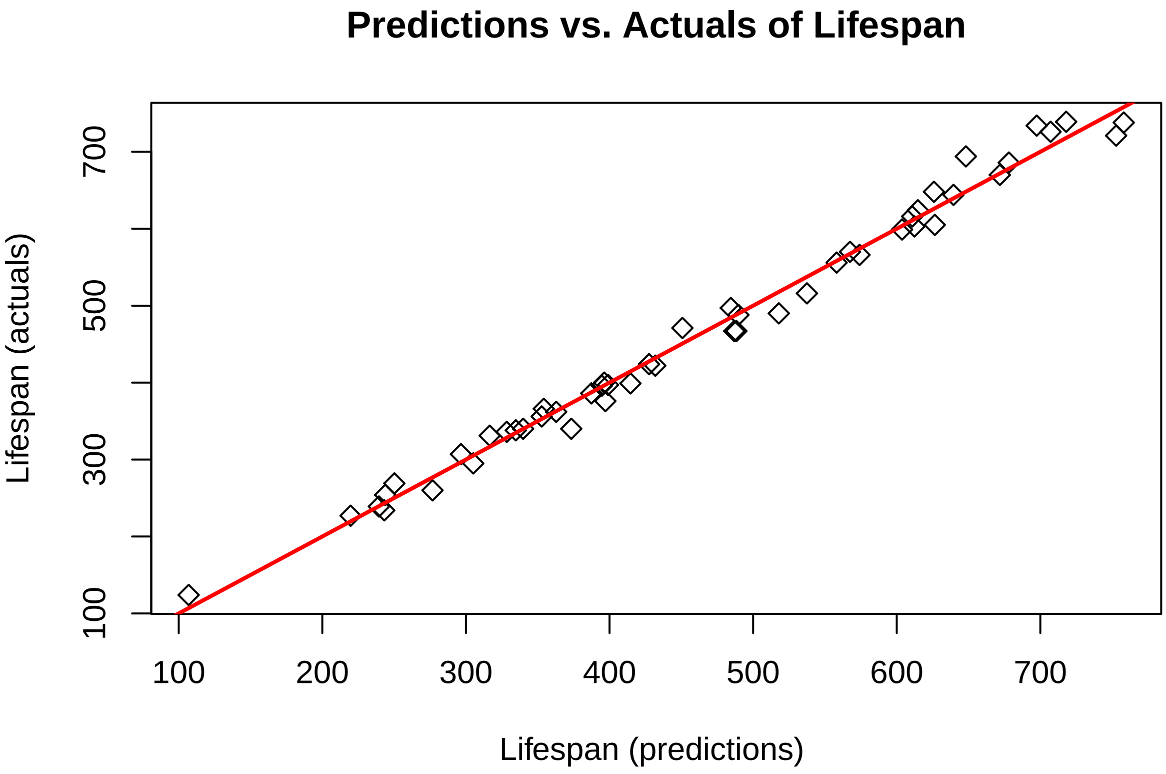


Figure 5: Plot of fitted values against actual lifespan

Presented below is the coefficient function of the model along with 95% confidence interval.

```
## [1] "done"
```

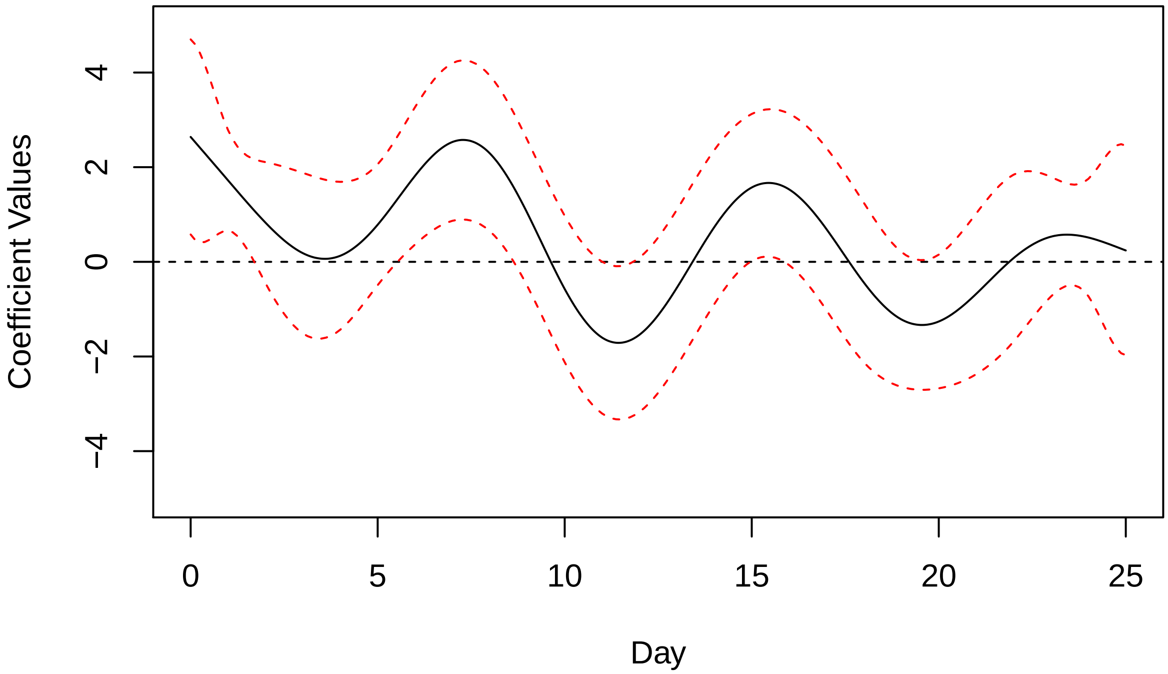


Figure 6: Coefficient Function of Functional Linear Regression Model

5 Question 4

5.1 Question

Conduct a permutation test for the significance of the regression. Calculate the R2 for your regression.

5.2 Solution

Below are the results of the permutation testing of the model to verify its significance over a confidence interval of 95%.

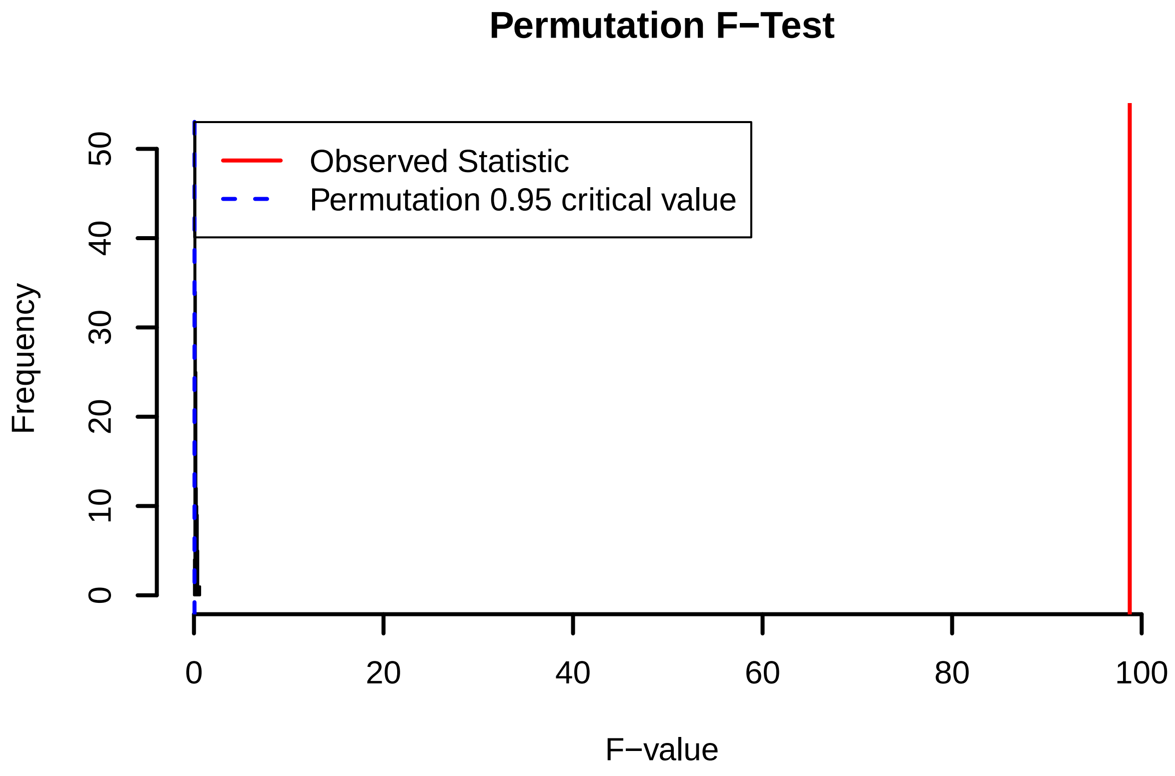


Figure 7: Permutation Testing Results

```
## The p-value is : 0
```

From the results of the P-value we can verify that indeed the functional regression model is significant.

Below is the R_squared (coefficient of determination) of the model.

```
## R_squared of regression model is:  0.9880245
```

6 Question 5

6.1 Question

Try a linear regression of lifespan on the principal component scores from your analysis. What is the R2 for this model? Does lm find that the model is significant? Reconstruct and plot the coefficient function for this model along with confidence intervals. How does it compare to the model obtained through functional linear regression?

6.2 Solution

In this section a linear regression model is set up using the scores of the principal components calculated earlier. Even though, it was estimated that the first 3 principal components of the dataset captured over 90% of the variability, the model created here uses all the principal components (10).

```
## 
## Call:
## lm(formula = lifetime ~ pca.scores)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -39.121  -9.727   0.088   8.468  48.630 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  466.26000    2.39889 194.365  < 2e-16 ***
## pca.scores1    2.14825    0.04019  53.459  < 2e-16 ***
## pca.scores2   -2.30641    0.05806 -39.726  < 2e-16 ***
## pca.scores3    0.28041    0.12279   2.284   0.0279 *  
## pca.scores4    1.20977    0.18159   6.662 6.26e-08 ***
## pca.scores5   -0.01300    0.45646  -0.028   0.9774    
## pca.scores6    1.13701    1.30351   0.872   0.3884    
## pca.scores7    5.57345    2.49907   2.230   0.0316 *  
## pca.scores8   -7.60772    4.91539  -1.548   0.1298    
## pca.scores9  -23.55840    9.10936  -2.586   0.0136 *  
## pca.scores10   8.44942   14.93138   0.566   0.5747    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 16.96 on 39 degrees of freedom
## Multiple R-squared:  0.9914, Adjusted R-squared:  0.9892 
## F-statistic: 451.2 on 10 and 39 DF,  p-value: < 2.2e-16
```

Below the results of the linear regression models of both the PCA scores and the functional linear regression are compared.

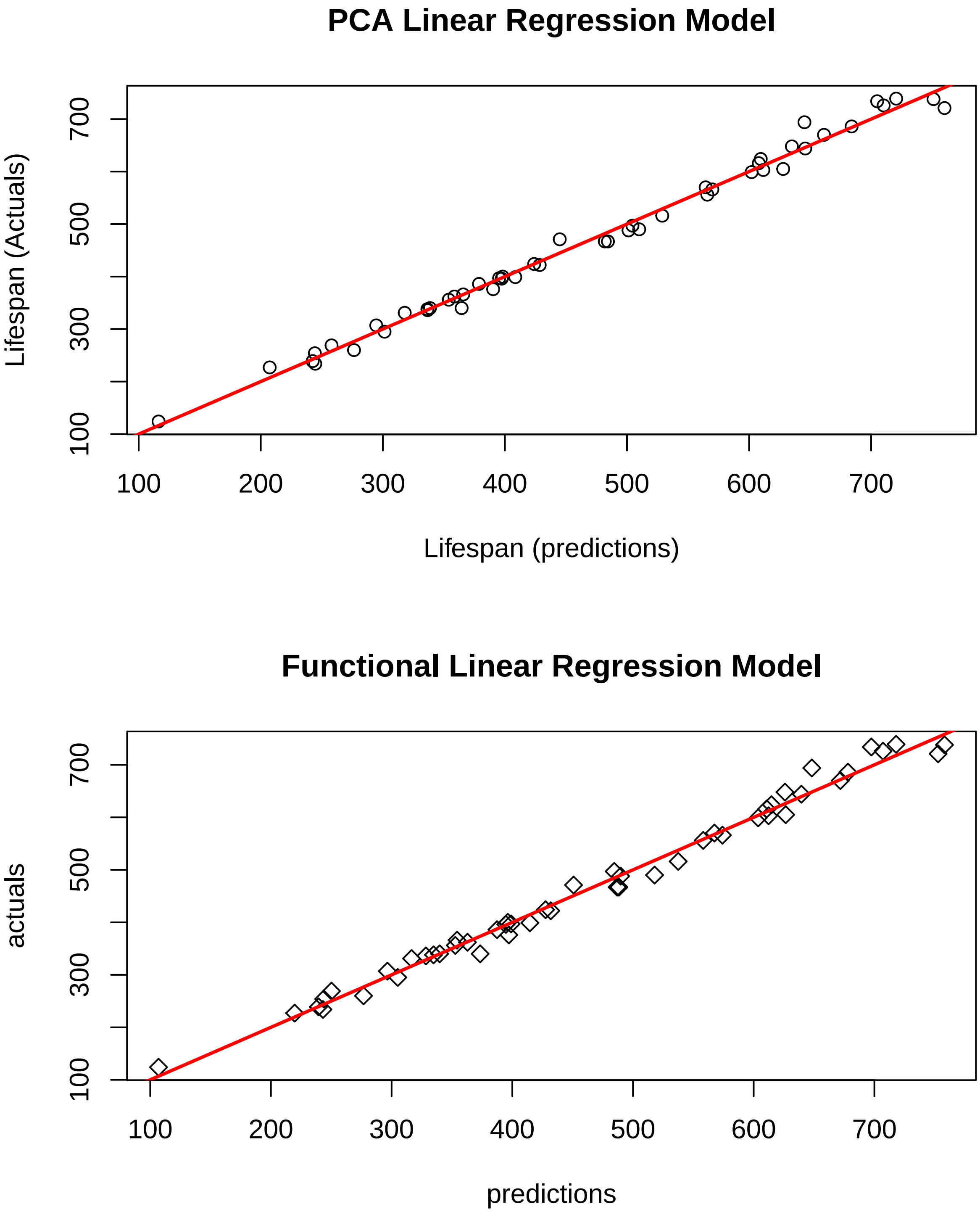
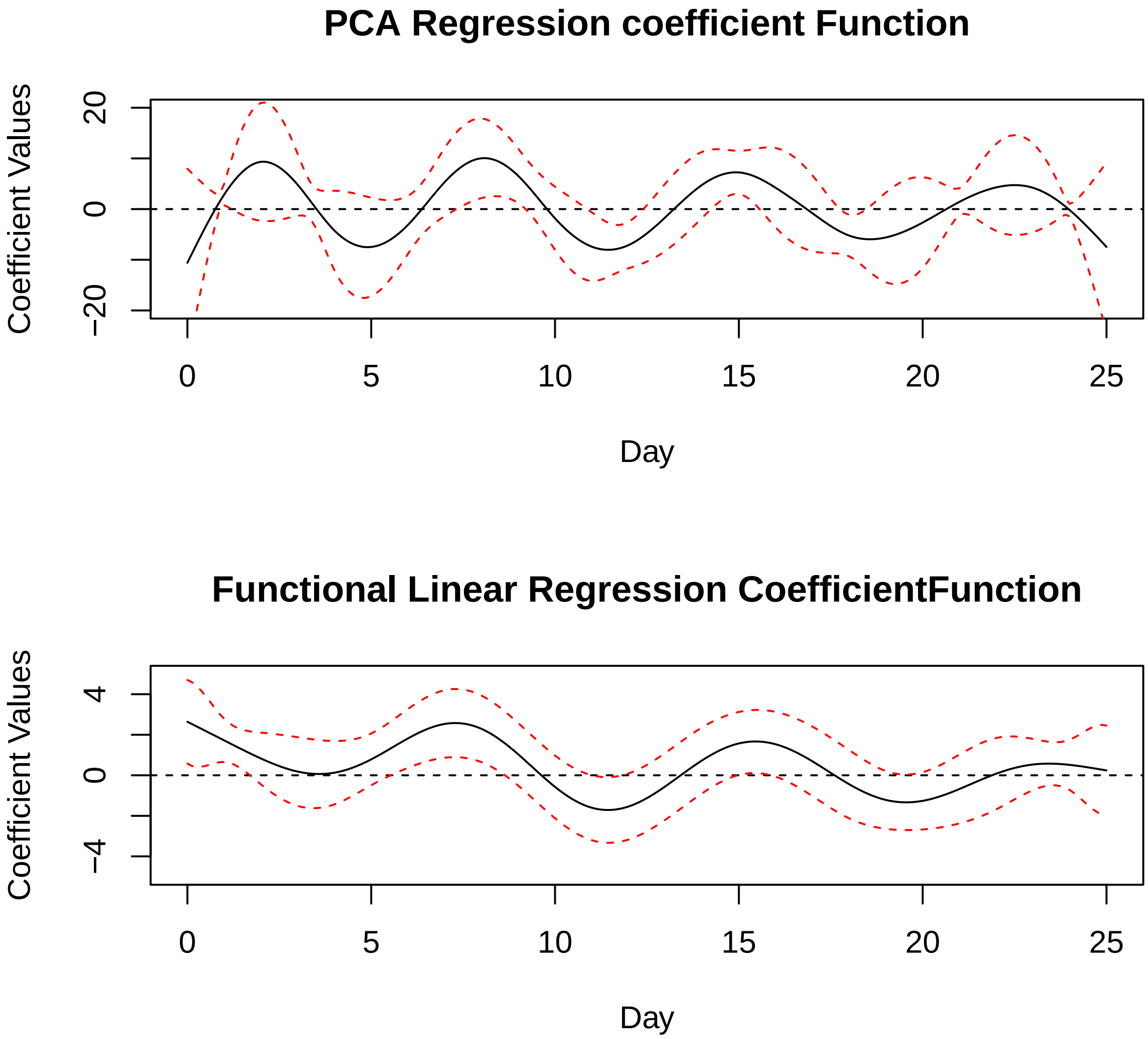


Figure 8: Comparison of PCA and Functional Regression Models Based on Fitted Values

From the above, it is evident that a similar trend and predictive performance is achieved by both models.

Below the coefficient function of the linear regression model of the principal components is reconstructed and compared with that of the functional linear regression model.

The confidence interval of both models are presented in red as well in the plots.


PCA Regression coefficient Function

Functional Linear Regression CoefficientFunction

Figure 9: Comparison of PCA and Functional Regression Models Based on Coefficient Functions

A similar trend of the coefficient function is observed in both models built from the principal components and using functional linear regression respectively.

7 References

1. Lecture Notes - Introduction to functional data analysis - Alejandra Cabana Nigro
2. http://faculty.bscb.cornell.edu/~hooker/FDA2008/