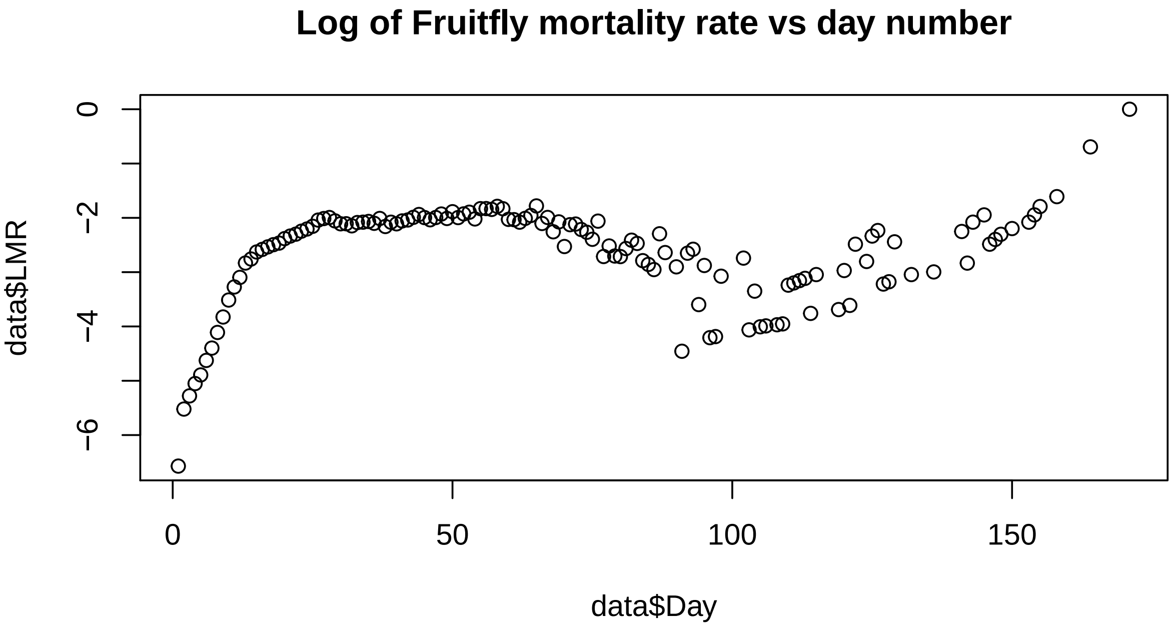
Computer Lab 5

Thomas Zhang

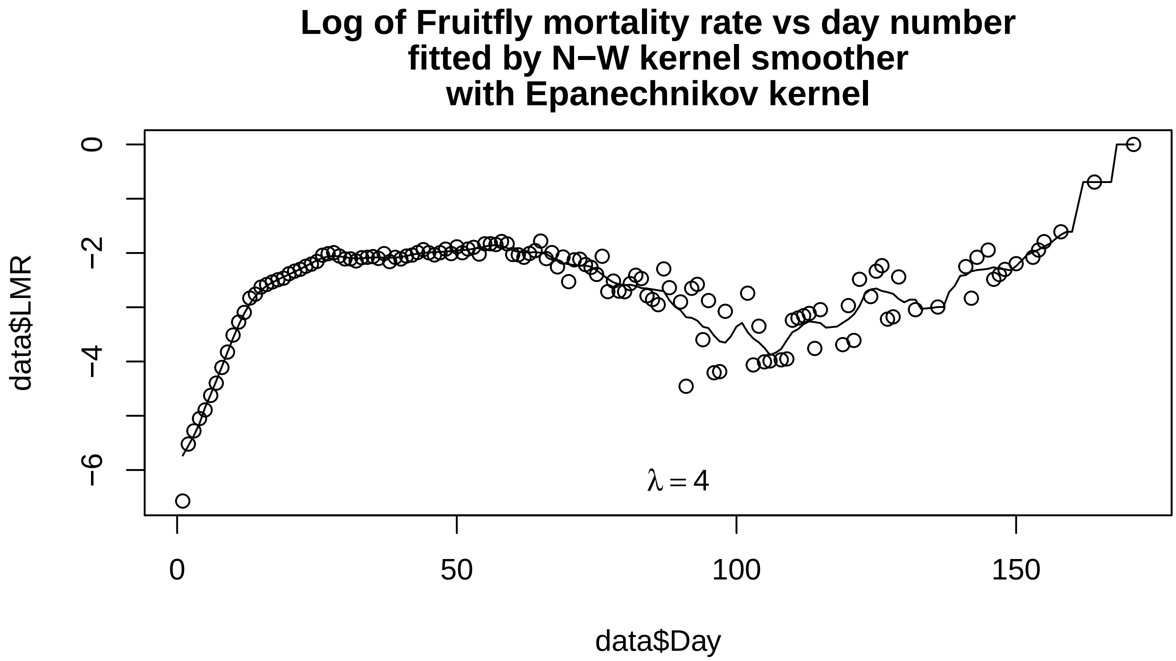
2015-11-22

Assignment 1

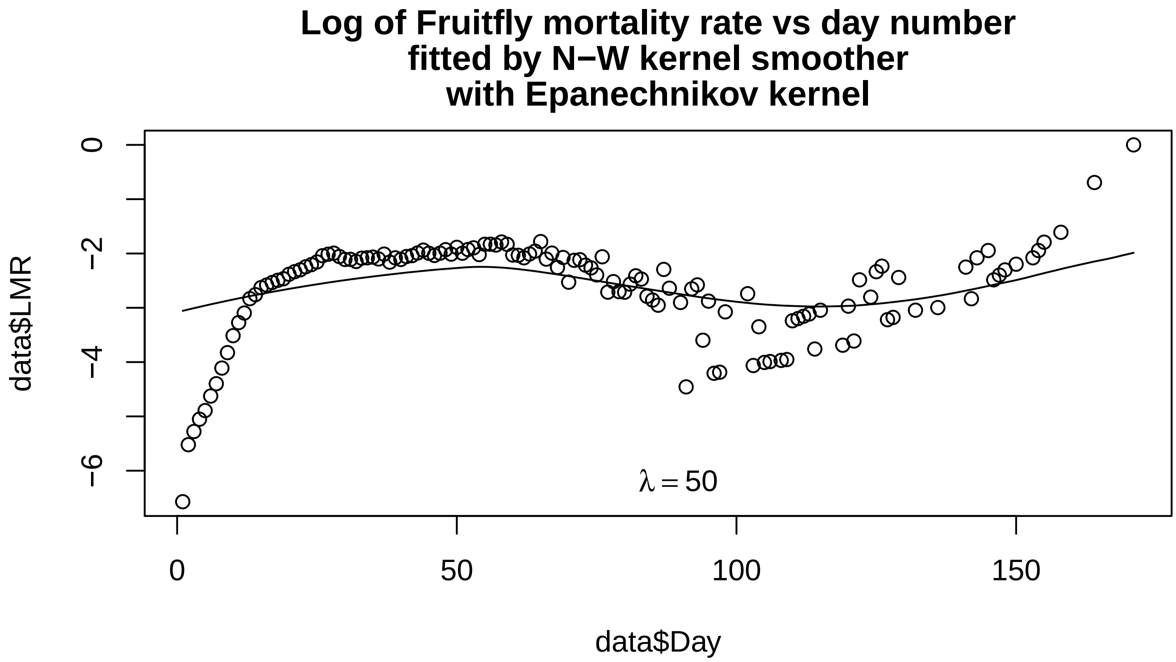
We plot the Log mortality rate (LMR) of fruit flies vs Day and see if there is a linear trend as suggested by Gompertz hypothesis.



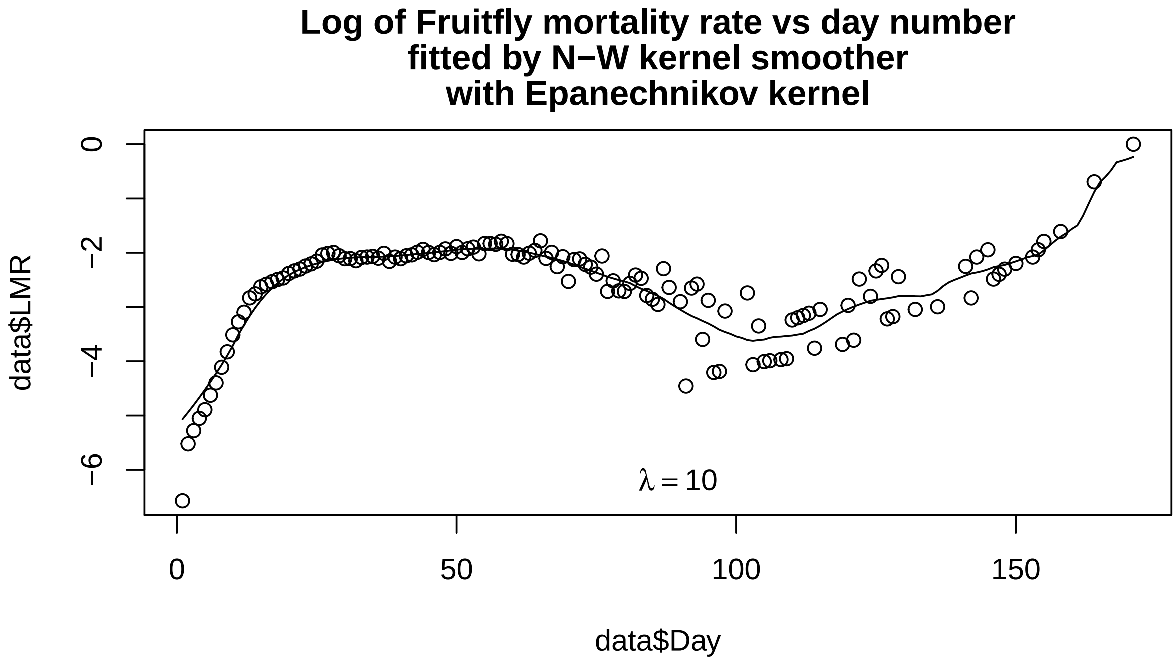
No linear trend is discernible. We proceed to fit a Nadaraya-Watson kernel smoother with Epanechnikov kernel and bandwidth parameter λ to the data. We vary the λ to produce different levels of "wigglyness" of the fitted curve.



Log of Fruitfly mortality rate vs day number
fitted by N−W kernel smoother
with Epanechnikov kernel



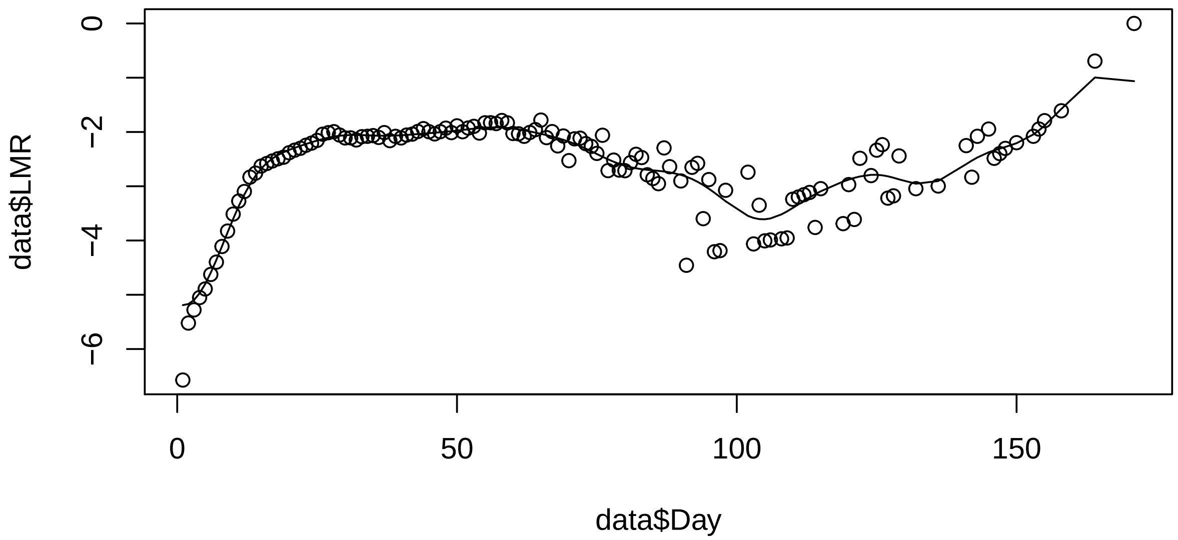
Log of Fruitfly mortality rate vs day number
fitted by N−W kernel smoother
with Epanechnikov kernel

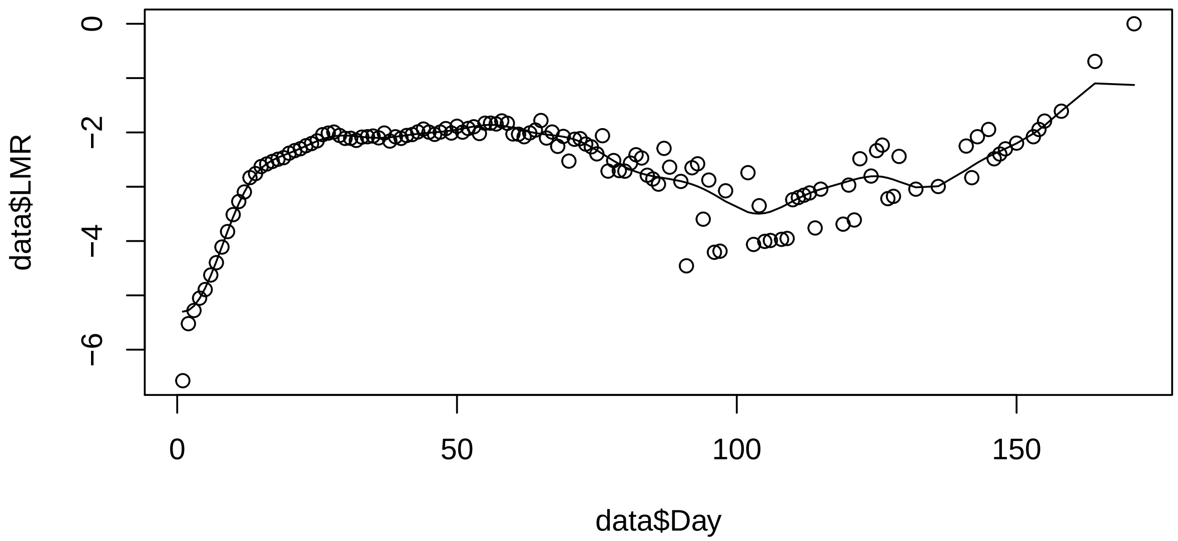


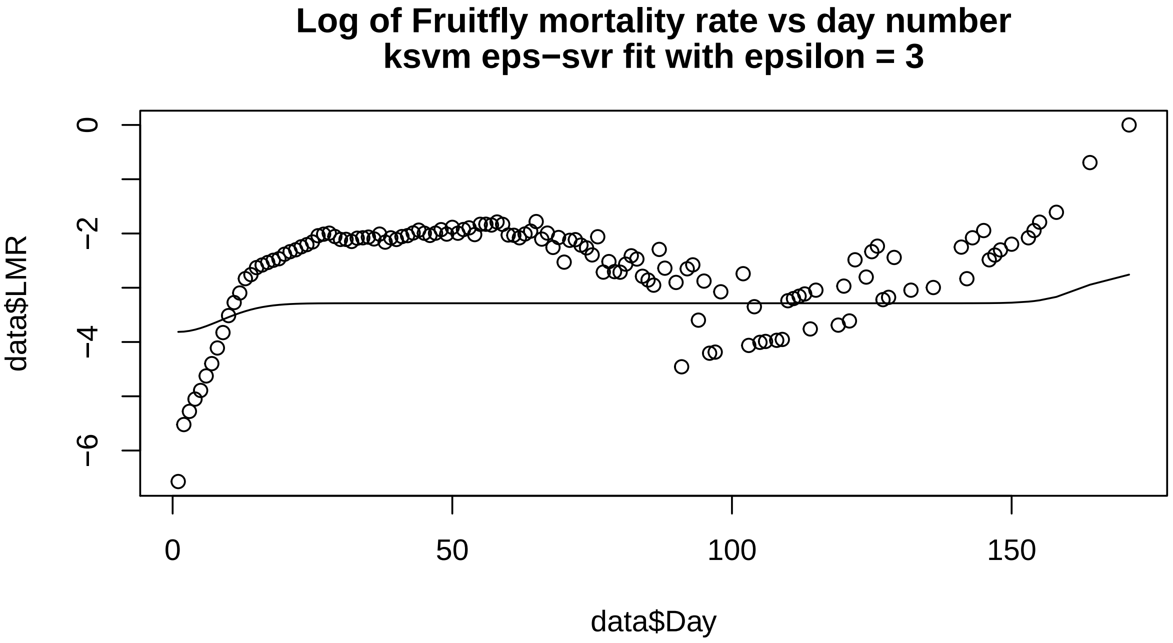
```
## [1] "MSE for training data for most reasonable model: 0.105"
```

From the plots, it can be inferred that the larger the number of λ, the smoother and flatter the fitted curve will be. This is not unexpected, as λ controls how fast the weights of the data points diminish with distance in Days. We see that the Nadaraya-Watson kernel smoother with λ=10 is the most reasonable fit and we have calculated the MSE for that fit.

We now try to fit the data using a SVM regression with RBF kernel and see how the fit changes with value on ε, where ε is the vertical distance the data points may be separated from the fitted curve before they start contributing towards the penalty term of the SVM regression.


Log of Fruitfly mortality rate vs day number
ksvm eps−svr fit with epsilon = 0.1


Log of Fruitfly mortality rate vs day number
ksvm eps−svr fit with epsilon = 0.005

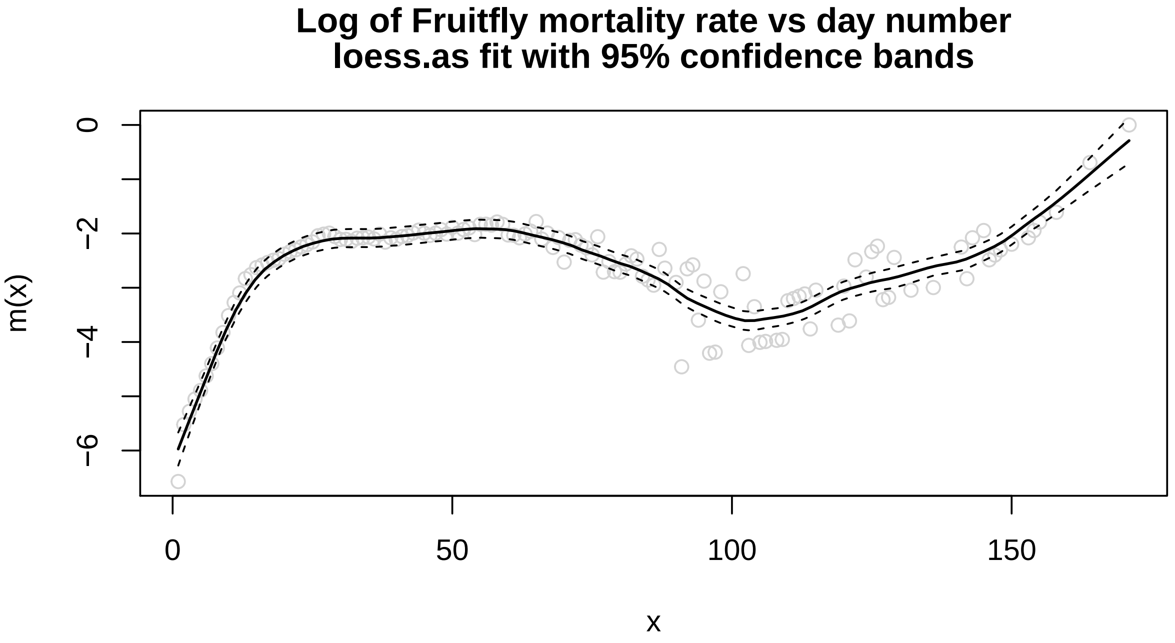


```
## [1] "MSE for training data for most reasonable ksvm eps-svr model: 0.131127"
```

We have chosen ε=0.005 as the most reasonable model and calculated its training data MSE. The most reasonable model looks similar to the most reasonable model of the Nadaraya-Watson kernel smoother fit.

It seems that when ε increases, the fit becomes more underfitted, as a result of not penalizing the data points inside the widened ε-insensitive tube around the fit. The number of support vectors also decreases, as according to theory only vectors outside the tube will non-zero terms in the kernel expansion expression.

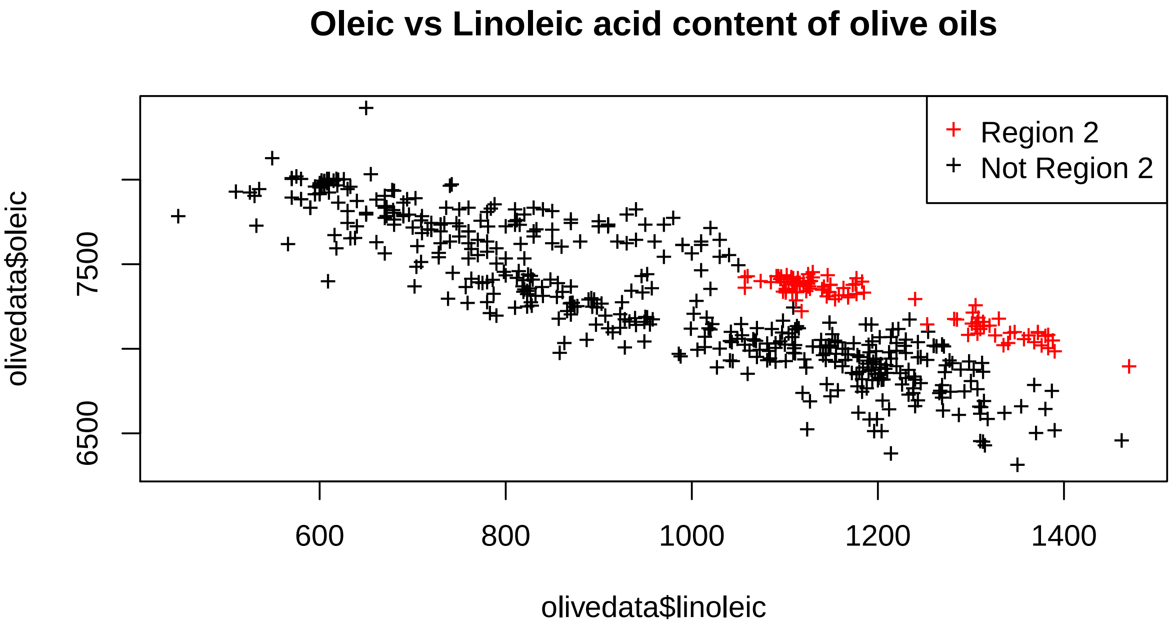
Now we use function `loess.as` in R package `fANCOVA` to fit "a local polynomial regression with automatic smoothing parameter selection . . . with generalized cross-validation" where the local polynomials have degree 1 to the data and plot the fit and its confidence bands.



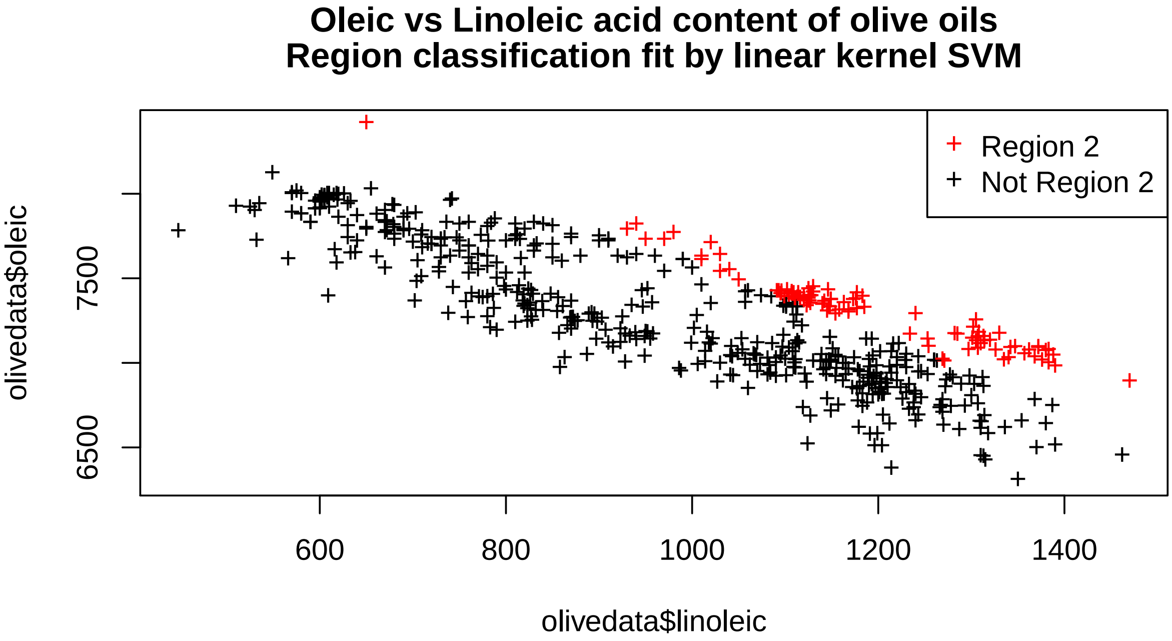
The fit looks rather nice. None of the plots however, based on this data set, support the Gompertz hypothesis which if true would produce a straight line.

Assignment 2

We plot the oleic vs linoleic acid contents of the olive oils coming from different regions of Italy, and color those olive oils coming from region 2 (entire Sardinia) in red.



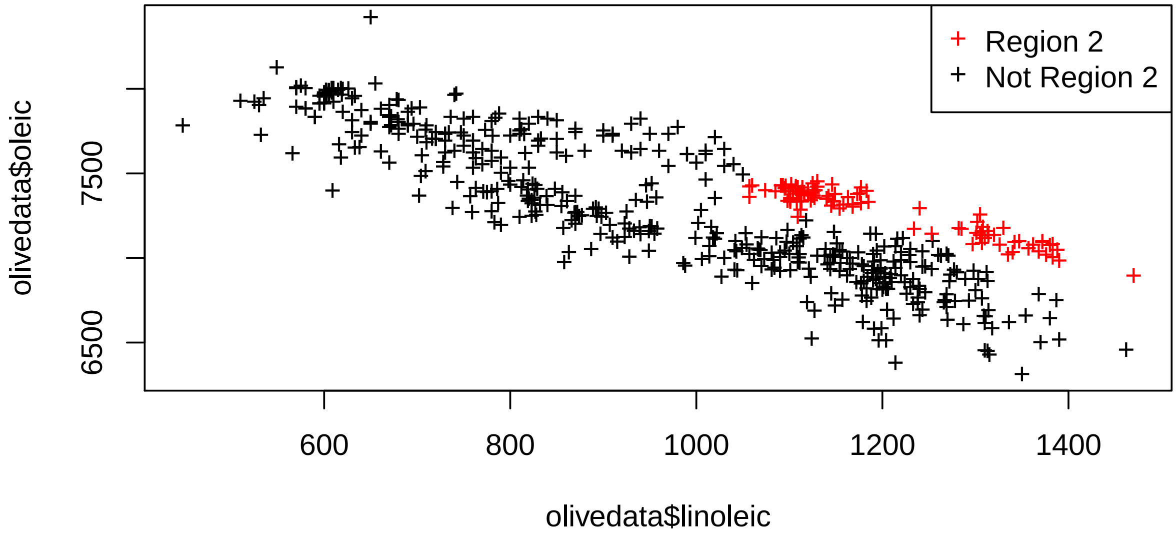
We see that it could be feasible to identify Region 2 oils based on oleic and linoleic acid contents, since they are clustered together in the scatter plot and that cluster does not contain oils from other Regions. We try to decide whether oils come from Region 2 or not using SVM, with linear kernel, gaussian RBF kernel,gaussian RBF kernel with increased penalty C and gaussian RBF kernel with increased bandwidth parameter σ respectively.



```
## [1] "Number of support vectors for linear kernel SVM C-svc: 119"
```

```
## [1] "Misclassification rate for linear kernel SVM: 0.0524"
```

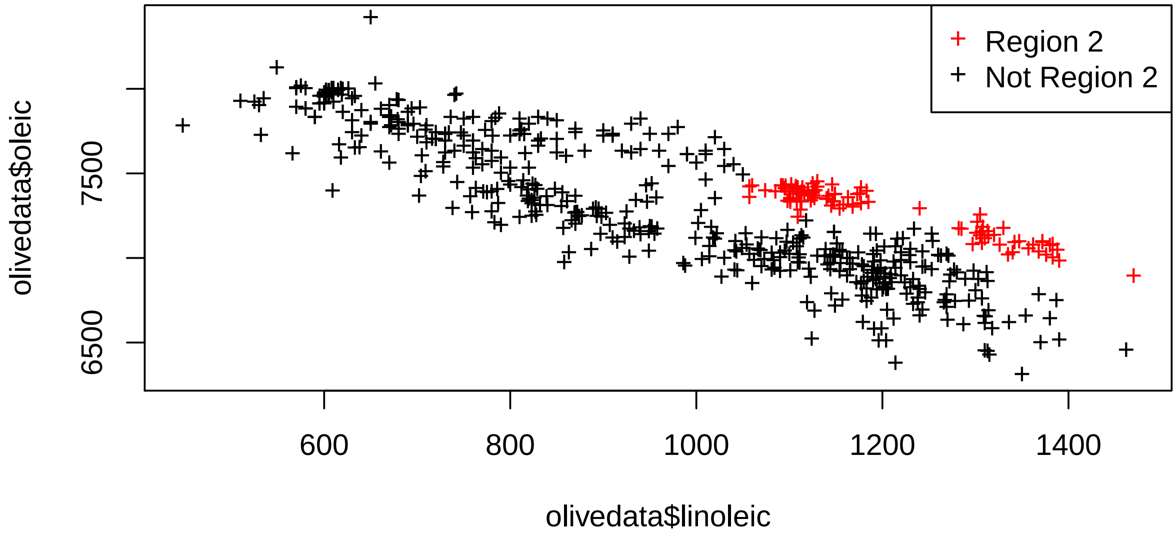
Oleic vs Linoleic acid content of olive oils
Region classification fit by Radial Basis kernel SVM
with sigma = 0.442 and C = 1



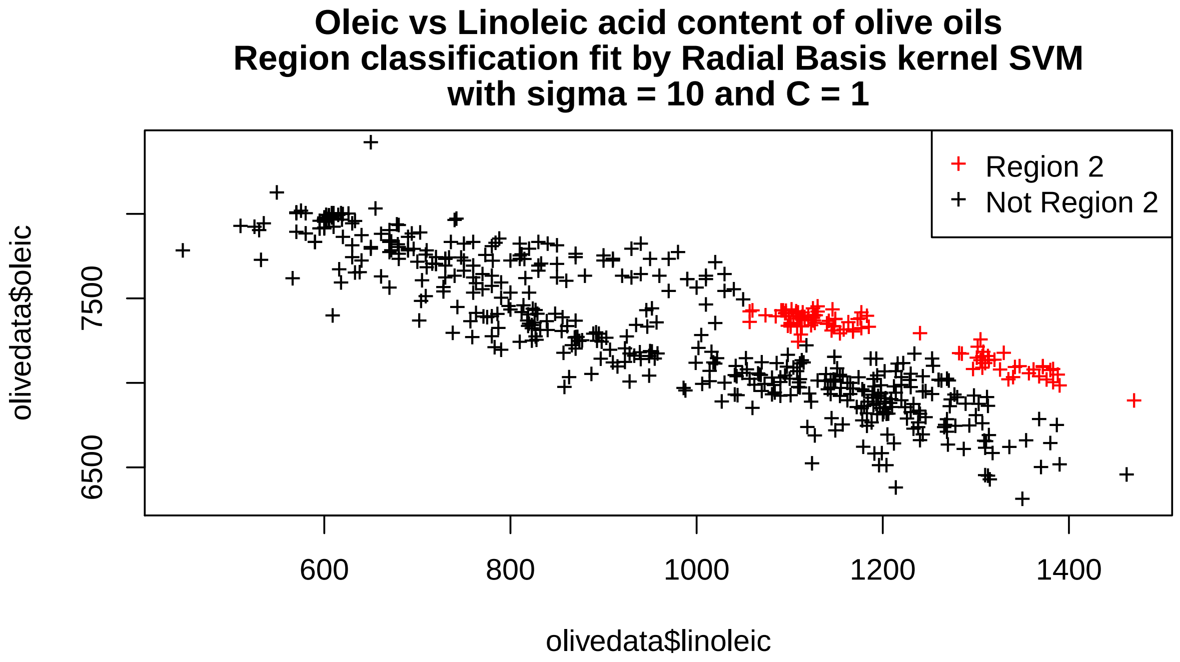
```
## [1] "Number of support vectors for Radial Basis kernel SVM C-svc: with sigma = 0.442 and C = 1: 52"
```

```
## [1] "Misclassification rate for Radial Basis kernel SVM: with sigma = 0.442 and C = 1: 0.00524"
```

Oleic vs Linoleic acid content of olive oils
Region classification fit by Radial Basis kernel SVM
with sigma = 0.442 and C = 100



```
## [1] "Number of support vectors for Radial Basis kernel SVM C-svc with sigma = 0.442 and C = 100: 32"
## [1] "Misclassification rate for Radial Basis kernel SVM with sigma = 0.442 and C = 100: 0.00524"
```



```
## [1] "Number of support vectors for Radial Basis kernel SVM C-svc with sigma = 10 and C = 1: 119"
```

```
## [1] "Misclassification rate for Radial Basis kernel SVM with sigma = 10 and C = 1: 0.00524"
```

We see that the three SVMs with gaussian RBF kernel performed better than the SVM with linear kernel where performance is measured by misclassification rate. this is not unexpected since linear kernal SVM can only construct linear decision boundries while gaussian RBF kernel SVM can construct non-linear boundries, such as the one better suited to separate region 2 oils from other oils. We also notice that when the penalty C is increased, the number of support vectors is decreased. This is because only misclassified data points or data points inside the SVM hyperplane margin can become support vectors. An increased value of C will decrease the size of the margin and make less vectors eligible to become support vectors. An increased value of gaussian kernel bandwidth parameter σ increases the number of support vectors because the kernel will have to consider vectors further from the data point being classified.

We also run a SVM multi-class classification with linear kernel that uses Region as response and all acids as predictors and estimate the 10-fold cross-validation score.

```
##  Setting default kernel parameters
```

```
## [1] "Number of support vectors : 53"
```

```
## [1] "Misclassification rate for 10-fold cross-validated : 0.0175"
```

```
## [1] "Cross validation score: 0.0298"
```

The model seems to work well.

Appendix - R-Code

```r
library(XLConnect)
library(fANCOVA)
library(kernlab)
#FROM THIS FILE LOCATION, EXCEL FILES SHOULD BE FOUND IN A SUBFOLDER IN THIS FILE LOCATION CALLED DATA
wb = loadWorkbook("D:/R_HW/ML-lab-1/data/mortality_rate.xls")
data = readWorksheet(wb, sheet = 1, header = TRUE)
LMR <- log(data$Rate)
data[,3] <- LMR
names(data)[3] <- "LMR"
#head(data)

epanechnikov <- function(x_0,x,lambda){
  if(abs(x_0-x)/lambda < 1){
    res <- 3/4 *  (1- 1/(lambda)^2 *(x_0 - x)^2)
  } else{
    res <- 0
  }
  return(res)
}

NW_ker_smooth <- function(X,Y,Xtest,lambda){
  kersum <- matrix(0,length(Xtest),length(X))
  for(i in 1:length(Xtest)){
    for(j in 1:length(X)){
      kercomp <- epanechnikov(Xtest[i],X[j],lambda)
      kersum[i,j] <- kercomp
    }
  }
  kersum_rowsums <- rowSums(kersum)
  #print(kersum_rowsums)
  pred <-c()
  for(i in 1:length(Xtest)){
    pred[i] <- (kersum[i,] %*% Y) /(kersum_rowsums[i])
  }
  return(pred)
}
Xtest <- seq(from=1,to=171, by=1)
plot(data$Day,data$LMR,main="Log of Fruitfly mortality rate vs day number")
#in order, wiggly (without producing NaNs),smooth and reasonable

sm1 <- NW_ker_smooth(data$Day,data$LMR,Xtest,4)
plot(data$Day,data$LMR,main=c("Log of Fruitfly mortality rate vs day number",
                              "fitted by N-W kernel smoother",
                              "with Epanechnikov kernel"))
lines(sm1)
legend("bottom",legend = expression(lambda == 4),bty="n")
sm3 <- NW_ker_smooth(data$Day,data$LMR,Xtest,50)
plot(data$Day,data$LMR,main=c("Log of Fruitfly mortality rate vs day number",
                              "fitted by N-W kernel smoother",
                              "with Epanechnikov kernel"))
lines(sm3)
```

```r
legend("bottom",legend = expression(lambda == 50),bty="n")
sm4 <- NW_ker_smooth(data$Day,data$LMR,Xtest,10)
plot(data$Day,data$LMR,main=c("Log of Fruitfly mortality rate vs day number",
                              "fitted by N-W kernel smoother",
                              "with Epanechnikov kernel"))
lines(sm4)
legend("bottom",legend = expression(lambda == 10),bty="n")
sm5 <- NW_ker_smooth(data$Day,data$LMR,data$Day,10)
bestmse <- sum((data$LMR-sm5)^2)/(length(data$LMR))
paste("MSE for training data for most reasonable model:",signif(bestmse,3))
set.seed(12345)
ksvmepsregr <- ksvm(LMR ~ Day,data=data,type="eps-svr",kernel="rbfdot",epsilon=0.1)
set.seed(12345)
ksvmepsregr1 <- ksvm(LMR ~ Day,data=data,type="eps-svr",kernel="rbfdot",epsilon=0.005)
set.seed(12345)
ksvmepsregr2 <- ksvm(LMR ~ Day,data=data,type="eps-svr",kernel="rbfdot",epsilon=3)
plot(data$Day,data$LMR,main=c("Log of Fruitfly mortality rate vs day number",
                              "ksvm eps-svr fit with epsilon = 0.1"))
lines(data$Day,predict(ksvmepsregr,newdata=data))
plot(data$Day,data$LMR,main=c("Log of Fruitfly mortality rate vs day number",
                              "ksvm eps-svr fit with epsilon = 0.005"))
lines(data$Day,predict(ksvmepsregr1,newdata=data))
plot(data$Day,data$LMR,main=c("Log of Fruitfly mortality rate vs day number",
                              "ksvm eps-svr fit with epsilon = 3"))
lines(data$Day,predict(ksvmepsregr2,newdata=data))
paste("MSE for training data for most reasonable ksvm eps-svr model:",signif(ksvmepsregr1@error))
loessfit <- loess.as(data$Day,data$LMR,criterion="gcv",plot=TRUE,
                     main=c("Log of Fruitfly mortality rate vs day number",
                            "loess.as fit with 95% confidence bands"))
pl <- predict(loessfit,se=TRUE)
lines(data$Day,predict(loessfit,data$Day)-2*pl$se.fit,lty=2)
lines(data$Day,predict(loessfit,data$Day)+2*pl$se.fit,lty=2)
library(kernlab)
olivedata <- read.csv("D:/R_HW/ML-lab-1/data/olive.csv")
R2 <- as.factor(olivedata$Region == 2)
olivedata[,12] <- R2
names(olivedata)[12] <- "R2"
n <- dim(olivedata)[1]
# head(olivedata)

plot(olivedata$linoleic, olivedata$oleic,pch="+",
     col=ifelse(olivedata$R2==TRUE, "red", "black"),
     main="Oleic vs Linoleic acid content of olive oils")
legend("topright",legend = c("Region 2","Not Region 2"),
       pch = "+",col=c("red","black"))
nSupVec <- function(ksvmobj){
  return(ksvmobj@nSV)
}

misclassrate <- function(truevec,retvec){
  counter <- 0
  for(i in 1:length(truevec)){
    if(truevec[i] != retvec[i]){
      counter <- counter + 1
    }
  }
  return(signif(counter/length(truevec),3))
}

set.seed(12345)
linksvm <-ksvm(R2~ oleic + linoleic,olivedata,kernel="vanilladot",kpar=list())
plot(olivedata$linoleic, olivedata$oleic,pch="+",
     col=ifelse(linksvm@fitted==TRUE, "red", "black"),
     main=c("Oleic vs Linoleic acid content of olive oils",
            "Region classification fit by linear kernel SVM"))
legend("topright",legend = c("Region 2","Not Region 2"),
       pch = "+",col=c("red","black"))
paste("Number of support vectors for linear kernel SVM C-svc:",nSupVec(linksvm))
paste("Misclassification rate for linear kernel SVM:",misclassrate(R2,linksvm@fitted))

set.seed(12345)
sigma <- signif(sigest(R2~ oleic + linoleic,olivedata)[2],3)

set.seed(12345)
rbfksvm <-ksvm(R2~ oleic + linoleic,olivedata,kernel="rbfdot")
plot(olivedata$linoleic, olivedata$oleic,pch="+",
     col=ifelse(rbfksvm@fitted==TRUE, "red", "black"),
     main=c("Oleic vs Linoleic acid content of olive oils",
            "Region classification fit by Radial Basis kernel SVM",
            paste("with sigma =",sigma,
                  "and C = 1")))
legend("topright",legend = c("Region 2","Not Region 2"),
       pch = "+",col=c("red","black"))
paste("Number of support vectors for Radial Basis kernel SVM C-svc:",
      "with sigma =",sigma,
      "and C = 1:",nSupVec(rbfksvm))
paste("Misclassification rate for Radial Basis kernel SVM:",
      "with sigma =",sigma,
      "and C = 1:",misclassrate(R2,rbfksvm@fitted))

set.seed(12345)
rbfksvm1 <-ksvm(R2~ oleic + linoleic,olivedata,kernel="rbfdot", C=100)
plot(olivedata$linoleic, olivedata$oleic,pch="+",
     col=ifelse(rbfksvm1@fitted==TRUE, "red", "black"),
     main=c("Oleic vs Linoleic acid content of olive oils",
            "Region classification fit by Radial Basis kernel SVM",
            paste("with sigma =",sigma,
                  "and C = 100")))
legend("topright",legend = c("Region 2","Not Region 2"),
       pch = "+",col=c("red","black"))
paste("Number of support vectors for Radial Basis kernel SVM C-svc",
      "with sigma =",sigma,
      "and C = 100:",nSupVec(rbfksvm1))
paste("Misclassification rate for Radial Basis kernel SVM",
      "with sigma =",sigma,
      "and C = 100:",misclassrate(R2,rbfksvm1@fitted))
```

```
set.seed(12345)
rbfksvm2 <-ksvm(R2~ oleic + linoleic,olivedata,kernel="rbfdot",kpar=list(sigma=10))
plot(olivedata$linoleic, olivedata$oleic,pch="+",
     col=ifelse(rbfksvm2@fitted==TRUE, "red", "black"),
     main=c("Oleic vs Linoleic acid content of olive oils",
            "Region classification fit by Radial Basis kernel SVM",
            paste("with sigma = 10 and C = 1")))
legend("topright",legend = c("Region 2","Not Region 2"),
       pch = "+",col=c("red","black"))
paste("Number of support vectors for Radial Basis kernel SVM C-svc with sigma = 10 and C = 1:",
      nSupVec(rbfksvm2))
paste("Misclassification rate for Radial Basis kernel SVM with sigma = 10 and C = 1:",
      misclassrate(R2,rbfksvm2@fitted))
olivedata2 <- cbind(olivedata[,2],olivedata[,4:11])
names(olivedata2)[1] <-"Region"
set.seed(12345)
allacidsksvm <- ksvm(Region~.,data=olivedata2,type="spoc-svc",kernel="vanilladot",cross=10)

paste("Number of support vectors :",
      nSupVec(allacidsksvm))
paste("Misclassification rate for 10-fold cross-validated :",
      misclassrate(olivedata2$Region,allacidsksvm@fitted))
paste("Cross validation score:",signif(allacidsksvm@cross,3))
## NA
```