STAT 360 - Handout #12: Segmented Regression

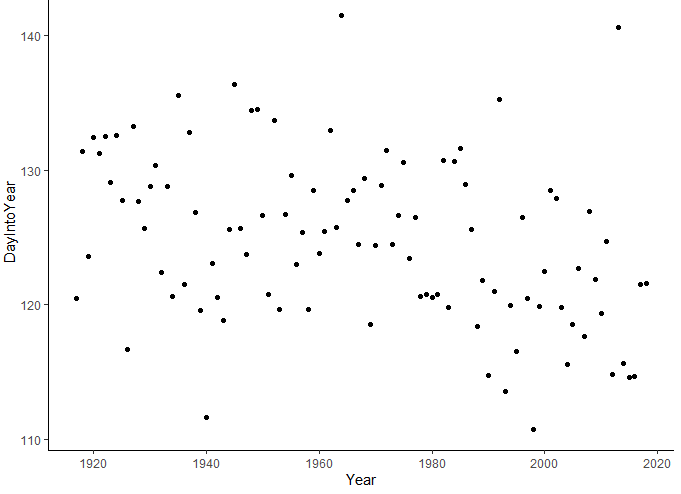
Example 12.1 For this handout, we will consider data from the Nenana River Ice Classic Competition. The goal of this lottery competition is to guess the exact time and date when the ice on the Nenana River breaks up.

Link: <https://www.nenanaakiceclassic.com/>

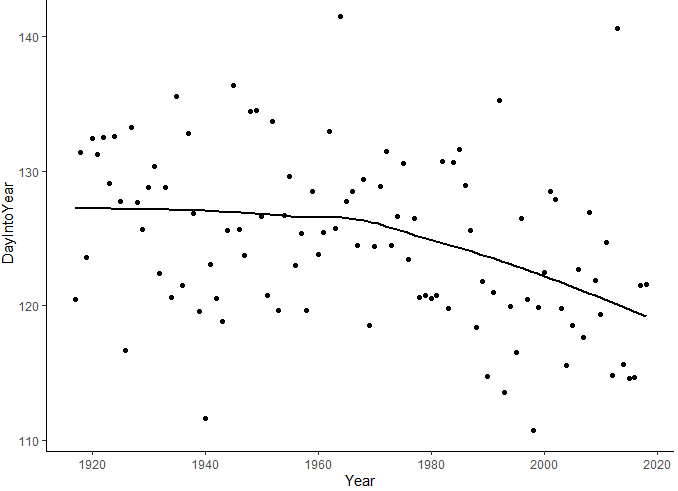
Note: This handout was created using data through 2018 – the updated data file includes data through 2020.

|  |  |
| --- | --- |
| Dataset |  |

A plot of the date (day of year) versus year.



Overall trend in this scatterplot…

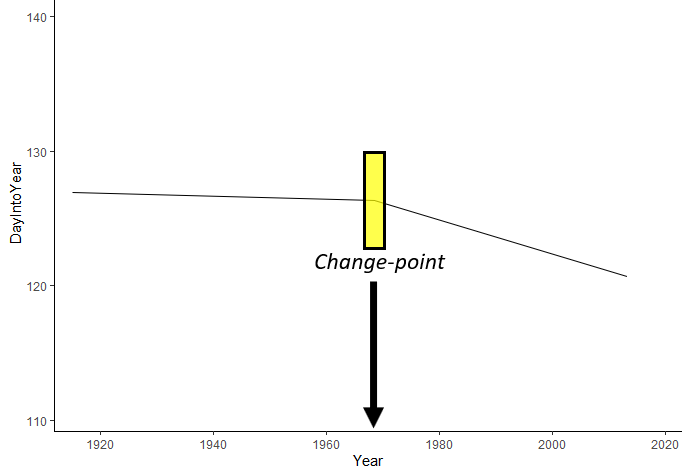


Questions

1. Describe the in the date of the ice break up from 1920 through the mid 1960’s?
2. How does this pattern change after the mid 1960’s?

Change-Point

A ***change-point*** is a point in which a structural change has occurred in the response variable. The structural change here appears to be a reduction in the expected date of ice break up, i.e. the ice break up date appears to be occurring earlier and earlier in recent years.



Model Setup

* Response Variable: DayofYear
* Predictor Variable: Year
* *Assume* the following structure for mean and variance functions

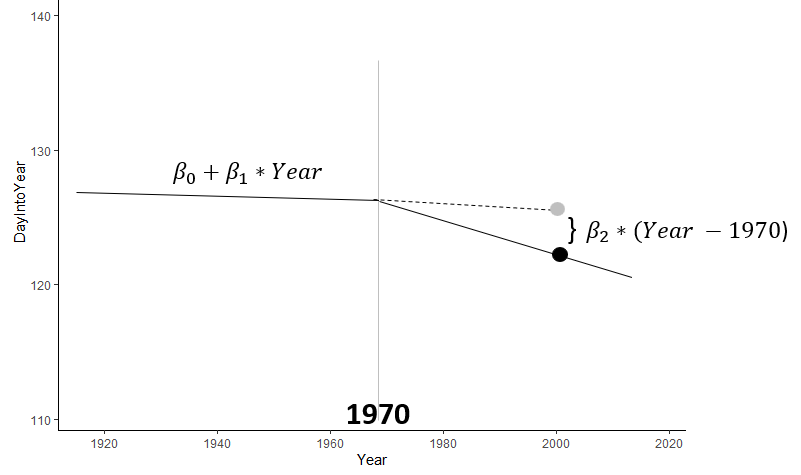
where

Understanding the Structure of the Segmented Regression Model

For the sake of discussion, suppose the change-point has been determined to be 1970.

1. What is the form of the model for years previous to 1970?
2. What is the form of the model for 1970?
3. What is the form of the model for years after 1970?
4. Write out the following model:

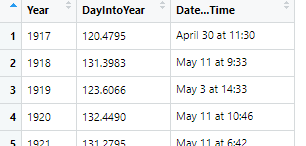
A visual depiction of the segmented regression model with a change-point at 1970.



Fitting this model in R

#Reading in the data

NenanaData <- read.csv("D:/Teaching/STAT360/Datasets/nenana.csv")



#Creating the new column for Distance\_Above\_ChangePoint

(NenanaData

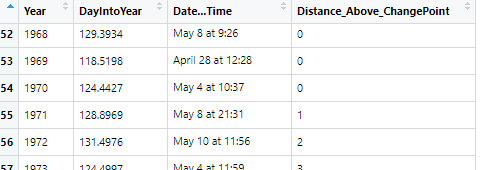
%>% mutate(Distance\_Above\_ChangePoint = case\_when( (Year - 1970) > 0 ~ (Year - 1970),

(Year - 1970) <= 0 ~ 0

)

)

)->NenanaData

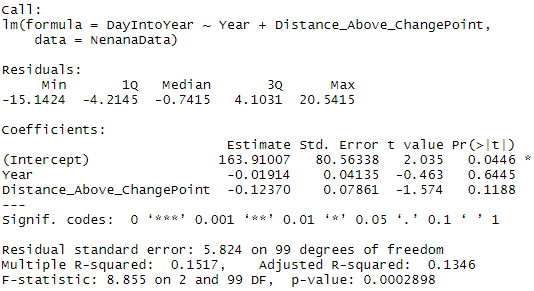


#Fitting the segmented regression model

NenanaData\_Fit <- lm(DayIntoYear ~ Year + Distance\_Above\_ChangePoint,

data=NenanaData)

summary(NenanaData\_Fit)



Questions

1. What is the R2 value of this model?

1. What information is gained by considering the p-value for the following test?

1. What information is gained by considering the p-value for the following test?

Plotting the residuals vs fitted and the segemented regression model in R

#Getting the predicted & residuals for this model

(NenanaData

%>% mutate(Predicted = predict(NenanaData\_Fit, newdata=NenanaData))

%>% mutate(Residuals = (DayIntoYear - Predicted))

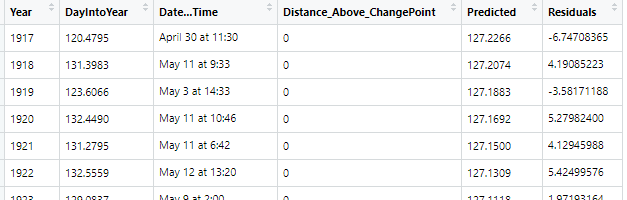
) -> NenanaData

#Getting the predicted values from this model

(NenanaData

%>% mutate(Predicted = predict(NenanaData\_Fit, newdata=NenanaData))

) -> NenanaData



#Plotting predicted with Change Point = 1970

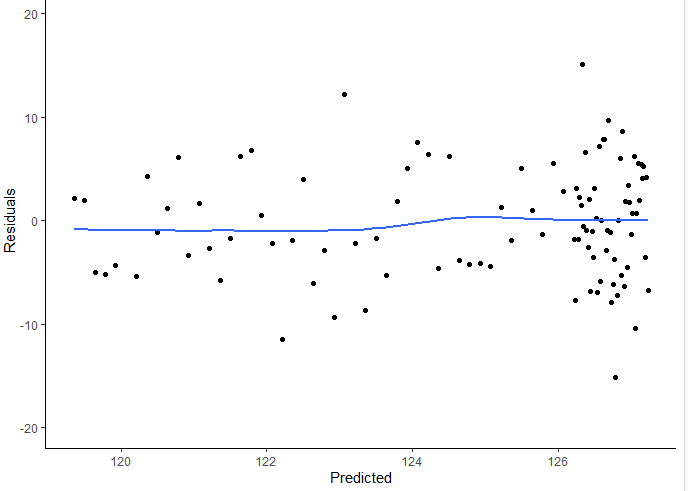
ggplot(NenanaData,mapping=aes(x=Predicted,y=Residuals)) +

geom\_point() +

geom\_smooth(method="loess",se=FALSE) +

ylim(-20,20) +

theme\_classic()



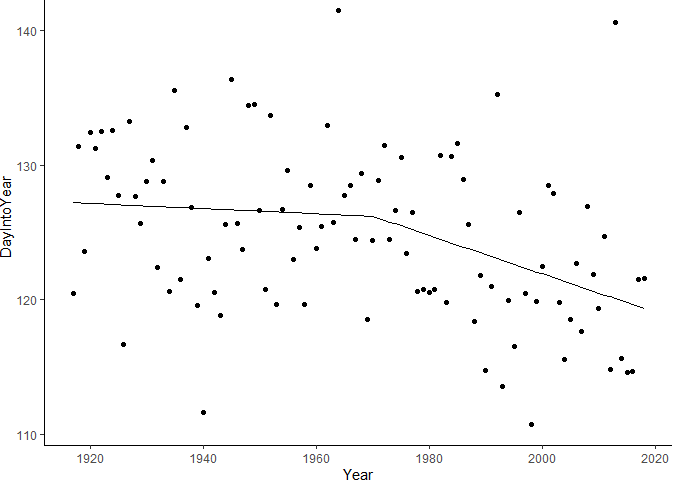
#Plotting predicted with Change Point = 1970

ggplot(NenanaData,mapping=aes(x=Year,y=DayIntoYear)) +

geom\_point() +

geom\_line(mapping=aes(x=Year,y=Predicted)) +

theme\_classic()



Rerun this model with Change Point = 1965

|  |  |
| --- | --- |
| Change Point | Residual Standard Error (RMSE) |
| 1970 | **5.824** |
| 1965 |  |

Question: Is the change point of 1965 appear to do better? Discuss

Finding the Optimal Change-Point

#Finding the optimal change-point

DataChangePoint <- data.frame(ChangePoint = seq(from=1960,to=1975,by=1))

(DataChangePoint

%>% rowwise()

%>% mutate (Sigma = (NenanaData

%>% mutate(Distance\_Above\_ChangePoint = case\_when( (Year - ChangePoint) > 0 ~ (Year - ChangePoint),

(Year - ChangePoint) <= 0 ~ 0

)

)

%>% lm(formula = DayIntoYear ~ Year + Distance\_Above\_ChangePoint)

%>% summary

%>% getElement("sigma")

))

) -> DataChangePoint

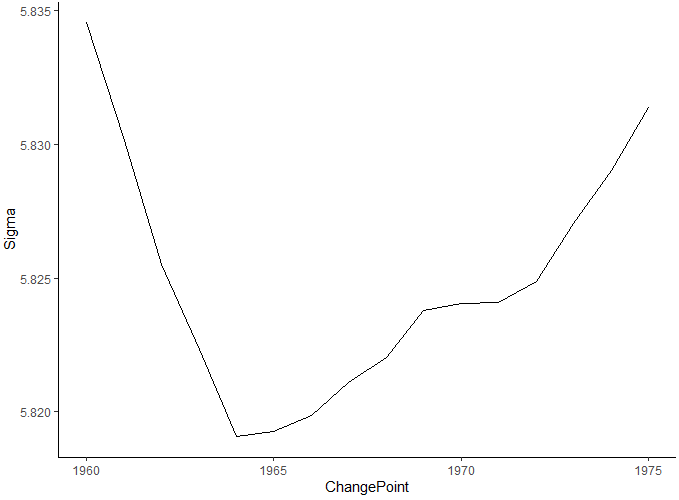
#Plotting the RMSE values for various change-points

library(ggplot2)

ggplot(DataChangePoint,mapping=aes(x=ChangePoint,y=Sigma))+

geom\_line()+

theme\_classic()



Questions:

1. What values appears to be the optimal change-point?
2. What is the practical interpretation of a model with a change-point = 1964?

Task: Rerun the regression model using a Change-Point = 1964. How well does this model do?

Fitting a Segmented Regression Model using the Segmented Package in R

#Load the library

library(segmented)

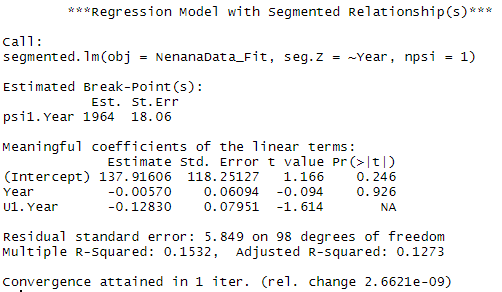
#Fit a SLR model with no change-point

NenanaData\_Fit <- lm(DayIntoYear ~ Year, data=NenanaData)

#Next, fit the segmented regression model with 1 change-point  
Segmented\_Fit <- segmented(NenanaData\_Fit,seg.Z=~Year, npsi=1)

#Summarise the fitted model

summary(Segmented\_Fit)



Example 12.2



|  |  |
| --- | --- |
|  |  |