Assignment of COPD Incidence in Scotland

by Fahmida Akter

## Load Dataset

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

data <- read\_csv("E:/DG Health Assignment/scotpho\_data\_extract\_COPD.csv")

## Rows: 32 Columns: 12

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): indicator, area\_name, area\_code, area\_type, period, definition, dat...  
## dbl (5): year, numerator, measure, lower\_confidence\_interval, upper\_confiden...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Data Preperation

#Variable list  
names(data)

## [1] "indicator" "area\_name"   
## [3] "area\_code" "area\_type"   
## [5] "year" "period"   
## [7] "numerator" "measure"   
## [9] "lower\_confidence\_interval" "upper\_confidence\_interval"  
## [11] "definition" "data\_source"

#head of data  
head(data)

## # A tibble: 6 x 12  
## indicator area\_name area\_code area\_type year period numerator measure  
## <chr> <chr> <chr> <chr> <dbl> <chr> <dbl> <dbl>  
## 1 Chronic obst~ Scotland S00000001 Scotland 2003 2002/03 t~ 6356. 187.  
## 2 Chronic obst~ Scotland S00000001 Scotland 2004 2003/04 t~ 6365 184.  
## 3 Chronic obst~ Scotland S00000001 Scotland 2005 2004/05 t~ 6322. 181.  
## 4 Chronic obst~ Scotland S00000001 Scotland 2006 2005/06 t~ 6362. 179.  
## 5 Chronic obst~ Scotland S00000001 Scotland 2007 2006/07 t~ 6796. 188.  
## 6 Chronic obst~ Scotland S00000001 Scotland 2008 2007/08 t~ 6675 182.  
## # ... with 4 more variables: lower\_confidence\_interval <dbl>,  
## # upper\_confidence\_interval <dbl>, definition <chr>, data\_source <chr>

#tail of data  
tail(data)

## # A tibble: 6 x 12  
## indicator area\_name area\_code area\_type year period numerator measure  
## <chr> <chr> <chr> <chr> <dbl> <chr> <dbl> <dbl>  
## 1 Chronic obst~ Scotland S00000001 Scotland 2013 2012/13 t~ 7210. 179   
## 2 Chronic obst~ Scotland S00000001 Scotland 2014 2013/14 t~ 7271. 178.  
## 3 Chronic obst~ Scotland S00000001 Scotland 2015 2014/15 t~ 7536. 181   
## 4 Chronic obst~ Scotland S00000001 Scotland 2016 2015/16 t~ 7675. 181.  
## 5 Chronic obst~ Scotland S00000001 Scotland 2017 2016/17 t~ 7493 174.  
## 6 Chronic obst~ Scotland S00000001 Scotland 2018 2017/18 t~ 7352 168.  
## # ... with 4 more variables: lower\_confidence\_interval <dbl>,  
## # upper\_confidence\_interval <dbl>, definition <chr>, data\_source <chr>

#dimension of data (variable and observation)  
dim(data)

## [1] 32 12

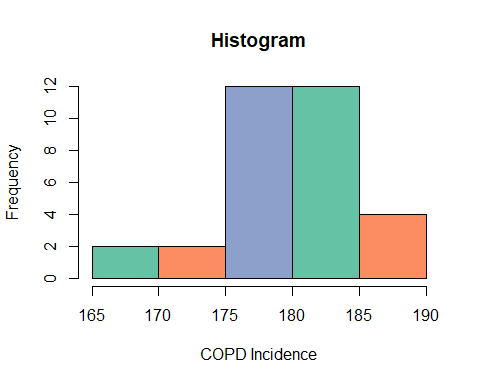
#checking of missing data  
mis <- is.na(data)  
sum(mis)

## [1] 0

## Data Visualization

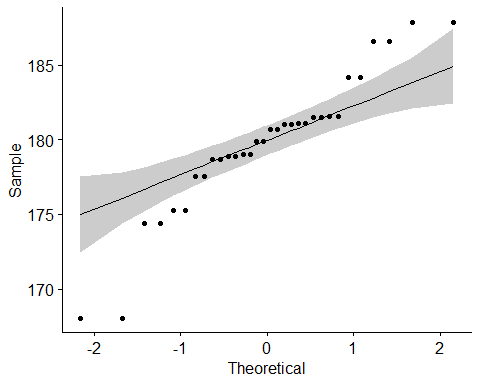
# Histogram

#Histogram  
library(RColorBrewer)  
hist(data$measure,breaks=3 ,col=brewer.pal(3,"Set2"),main="Histogram", xlab="COPD Incidence")

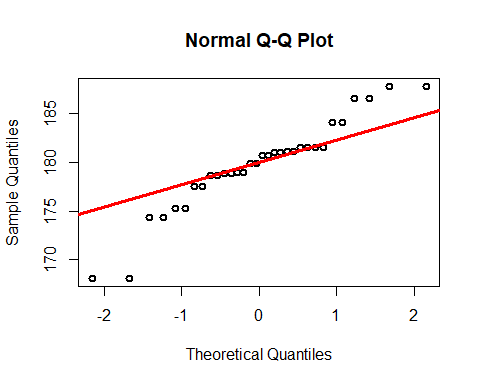


# Q-Q Plot

#Q-Q plot   
#install.packages("ggpubr")  
library(ggpubr)  
ggqqplot(data$measure)



#Q-Q Plot   
qqnorm(data$measure, pch =1, frame = TRUE, lwd=2)  
qqline(data$measure, col = "red", lwd = 3)



## Normality Test by Shapiro-Wilk normality test

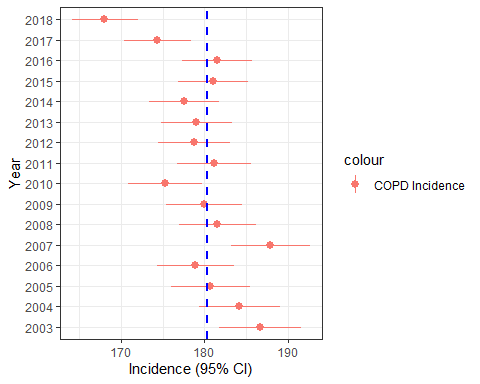
#Shapiro-Wilk normality test  
norm.test <- shapiro.test(data$measure)  
norm.test

##   
## Shapiro-Wilk normality test  
##   
## data: data$measure  
## W = 0.92952, p-value = 0.03797

From the output, the p-value < 0.05 implying that the distribution of the data are significantly different from normal distribution. In other words, we can not assume the normality.

## Forest Plot

label <- as.factor (c(2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012 ,2013, 2014, 2015, 2016, 2017, 2018, 2003, 2004, 2005, 2006, 2007, 2008 ,2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018))  
incidence <- data$measure  
lower <- data$lower\_confidence\_interval  
upper <- data$upper\_confidence\_interval  
  
#Data frame  
df <- data.frame(label, incidence, lower,upper)  
  
library(ggplot2)  
fp <- ggplot(data=df, aes(x=label, y=incidence, ymin=lower, ymax=upper, col="COPD Incidence")  
 ) +  
 geom\_pointrange() +   
 geom\_hline(yintercept=180.3, lty=2, col="blue", lwd=1) + # add a dotted line at x=1 after flip  
 coord\_flip() + # flip coordinates (puts labels on y axis)  
 xlab("Year") + ylab("Incidence (95% CI)") +  
 theme\_bw() # use a white background  
print(fp)



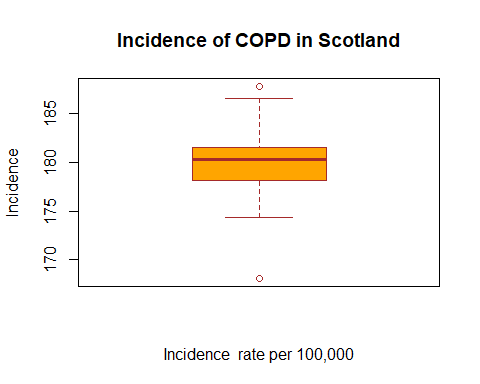
ggsave(path="E:/DG Health Assignment/", "forest plot.png", width = 20, height = 20, units = "cm", dpi=300)

## Descriptive Statistical Analysis

#Descriptive statistics of Incidence  
summary(data$measure)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 168.1 178.4 180.3 179.8 181.5 187.8

#Box plot of incidence  
boxplot(data$measure,  
main = "Incidence of COPD in Scotland"  
, xlab = "Incidence rate per 100,000", ylab = "Incidence",col = "orange", border = "brown", horizontal = FALSE, notch = FALSE)



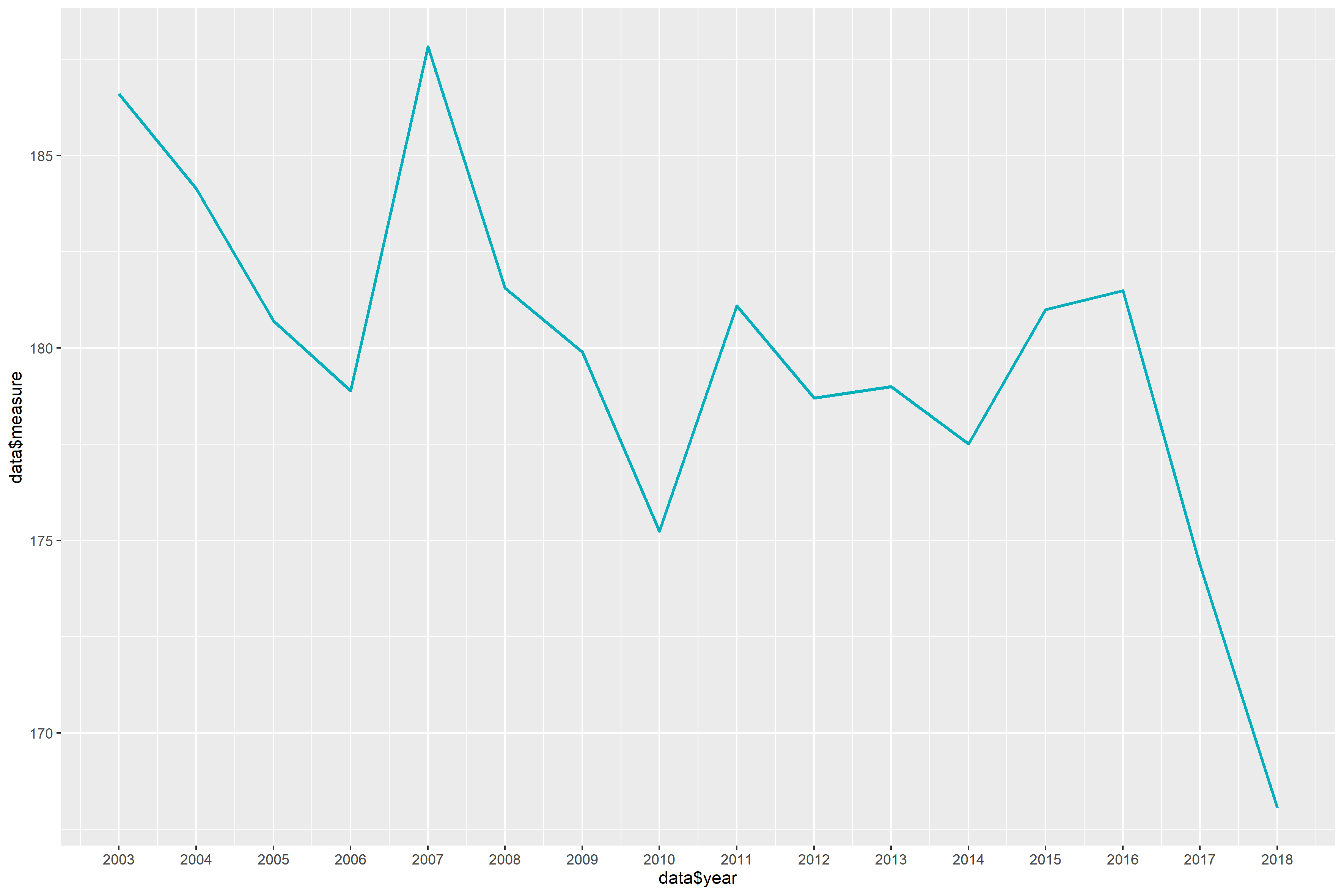
## Linear Model for trend analysis

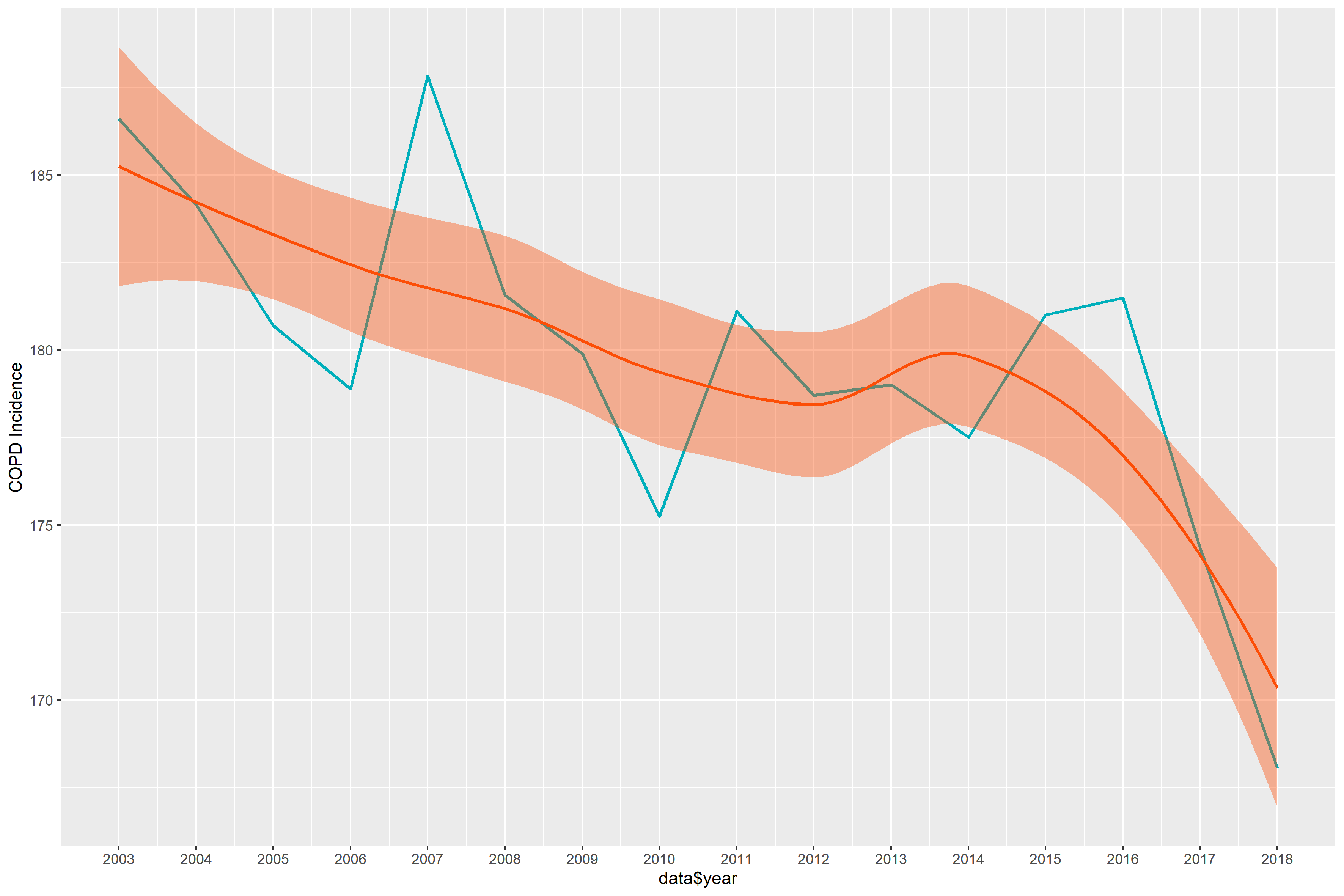
fit <- lm(data$measure~data$year)  
summary(fit)

##   
## Call:  
## lm(formula = data$measure ~ data$year)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.6174 -1.4401 0.0495 1.7052 5.7064   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1540.7150 266.9049 5.773 2.64e-06 \*\*\*  
## data$year -0.6769 0.1328 -5.099 1.76e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.462 on 30 degrees of freedom  
## Multiple R-squared: 0.4643, Adjusted R-squared: 0.4464   
## F-statistic: 26 on 1 and 30 DF, p-value: 1.761e-05

## Trend Analysis graph

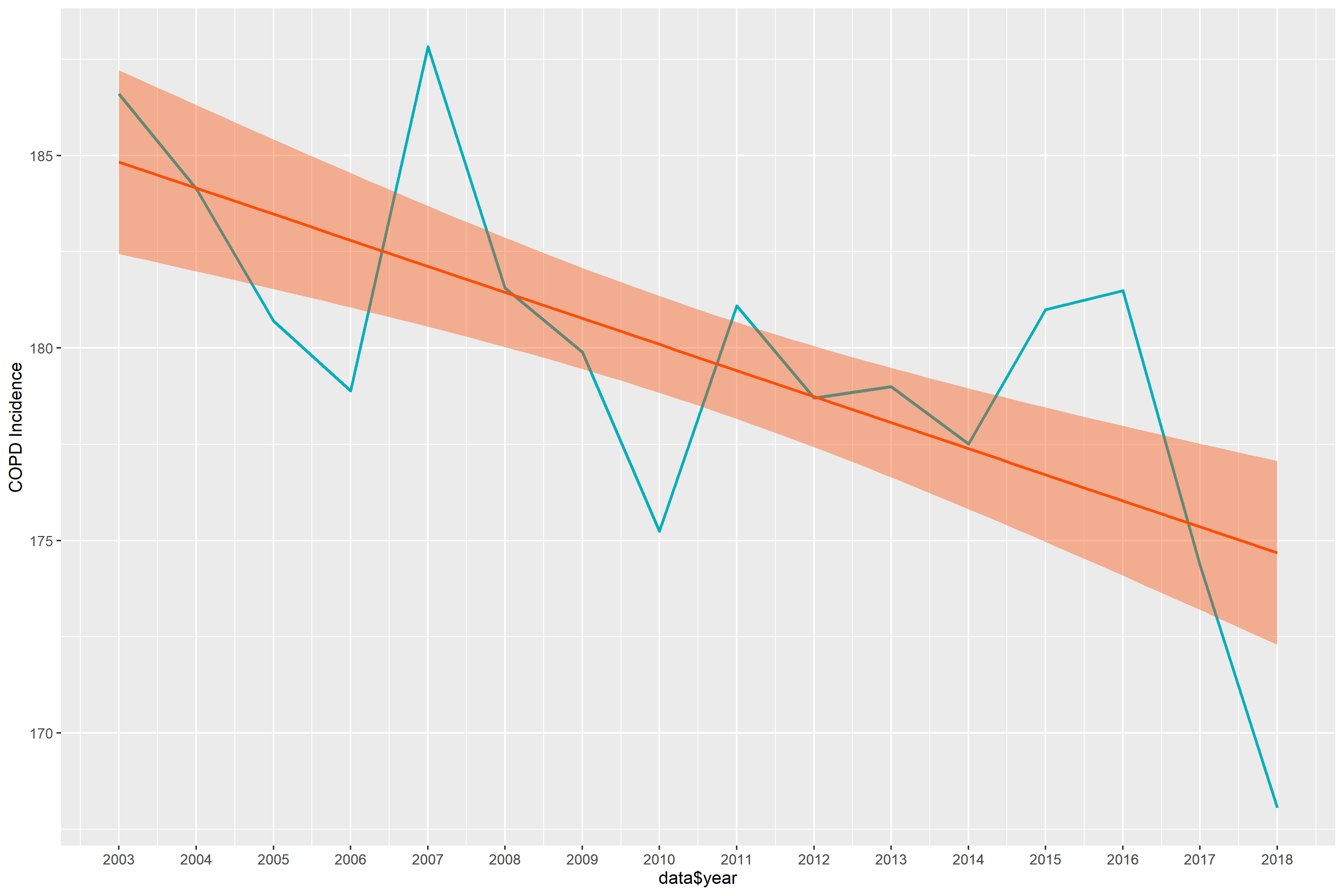
#line plot  
p <- ggplot(data = data, aes(x = data$year, y = data$measure)) +   
 geom\_line(color = "#00AFBB", size = 1) +  
 scale\_x\_continuous("data$year", labels = as.character(data$year), breaks = data$year)  
ggsave(path="E:/DG Health Assignment/", "line plot1.png", width = 30, height = 20, units = "cm", dpi=300)  
print(p)



# Polynomial Regression Fitting  
p + stat\_smooth(color = "#FC4E07", fill = "#FC4E07",  
 method = "loess") + xlab("Year") + ylab("COPD Incidence")   
 ggsave(path="E:/DG Health Assignment/", "line plot2.png", width = 30, height = 20, units = "cm", dpi=300)  


# Linear Model Fitting  
p + stat\_smooth(color = "#FC4E07", fill = "#FC4E07",  
 method = "lm") + xlab("Year") + ylab("COPD Incidence")

## `geom\_smooth()` using formula 'y ~ x'



ggsave(path="E:/DG Health Assignment/", "line plot3.png", width = 30, height = 20, units = "cm", dpi=300)

## `geom\_smooth()` using formula 'y ~ x'