# ISPPD Workshop #2 Evaluating Vaccine Impact using Time Series Data

Daniel M. Weinberger (daniel.weinberger@yale.edu) and Kayoko Shioda (kayoko.shioda@yale.edu)

Sunday, April 15, 2018

## Set up

First, download "Brazil_acp.csv" and save it in your folder.

Then, run the following section to import the dataset.

```
# Set working directory
setwd("C:/Users/dmw63/Dropbox (Personal)/ISPPD workshop") # Please update
this line

# Import the data in a .csv file
d <- read.csv("Brazil_acp.csv")
```

Let's explore the dataset a little bit...

```
# Explore the dataset
names(d)

##  [1] "age_group"        "date"             "J12_18"          
##  [4] "A10_B99_nopneumo" "A17"              "A18"             
##  [7] "A19"              "A39"              "A41"             
## [10] "B20_24"           "B34"              "B96"             
## [13] "B97"              "B99"              "C00_D48"         
## [16] "D50_89"           "E00_99"           "E10_14"          
## [19] "E40_46"           "G00_99_SY"        "H00_99_SY"       
## [22] "I00_99"           "I60_64"           "cJ20_J22"        
## [25] "K00_99"           "K35"              "K80"             
## [28] "L00_99"           "M00_99"           "N00_99"          
## [31] "N39"              "P00_99"           "P05_07"          
## [34] "Q00_99"           "S00_T99"          "U00_99"          
## [37] "V00_Y99"          "Z00_99"           "ACH_NOJ"

head(d)

##   age_group     date J12_18 A10_B99_nopneumo A17 A18 A19 A39 A41 B20_24
## 1       80+ 1/1/2004   3192             1357  NA  NA  NA  NA 249     NA
## 2       80+ 2/1/2004   3691             1389  NA  NA  NA  NA 275     NA
## 3       80+ 3/1/2004   6131             1604  NA  NA  NA  NA 305     NA
## 4       80+ 4/1/2004   5044             1377  NA  NA  NA  NA 258     NA
## 5       80+ 5/1/2004   4694             1385  NA  NA  NA  NA 260     NA
## 6       80+ 6/1/2004   4986             1449  NA  NA  NA  NA 295     NA
##   B34 B96 B97 B99 C00_D48 D50_89 E00_99 E10_14 E40_46 G00_99_SY H00_99_SY
## 1  NA  NA  NA  NA    1715    481   2349    866    800       696       271
## 2  NA  NA  NA  NA    1618    406   2221    770    780       668       280
## 3  NA  NA  NA  NA    2129    490   2393    828    805       672       374
## 4  NA  NA  NA  NA    1819    426   2123    727    767       624       279
## 5  NA  NA  NA  NA    1943    420   2178    796    796       754       355
## 6  NA  NA  NA  NA    1821    404   2131    773    745       729       330
##   I00_99 I60_64 cJ20_J22 K00_99 K35 K80 L00_99 M00_99 N00_99 N39 P00_99
## 1  12168   2683        0   2930   9 110    522    844   2070 227      1
## 2  11274   2511        0   2779  21 122    515    757   1896 220      1
## 3  12445   2592        1   3161  17 150    648    899   2224 229      3
## 4  11500   2594        2   2762  16 110    525    792   1915 205      0
## 5  11872   2731        2   2994  15 114    539    899   2021 232      0
## 6  12580   2764        3   2793  20 139    535    909   1898 234      1
##   P05_07 Q00_99 S00_T99 U00_99 V00_Y99 Z00_99 ACH_NOJ
## 1     NA     96    2016     NA      NA    190   30727
## 2     NA     69    1907     NA      NA    157   29844
## 3     NA     79    2076     NA      NA    215   33020
## 4     NA     74    2020     NA      NA    210   28916
## 5     NA     83    2402     NA      NA    177   30341
## 6     NA     71    2369     NA      NA    184   30565
```

```
table(d$age_group)
```

```
## 
##  <1 80+ 
## 120 120
```

Let's take a look at a date variable. How does it look like? Is it in a right format?

```
class(d$date) # "factor" --> Need to change it to "date"
```

```
## [1] "factor"
```

```
head(d$date)
```

```
## [1] 1/1/2004 2/1/2004 3/1/2004 4/1/2004 5/1/2004 6/1/2004
## 120 Levels: 1/1/2004 1/1/2005 1/1/2006 1/1/2007 1/1/2008 ... 9/1/2013
```

```
# Change the type of the date variable so that R can recognize it as a date
variable
d$date <- as.Date(d$date,format="%m/%d/%Y")
class(d$date) # Now it's changed to "Date"
```

```
## [1] "Date"
```

```
head(d$date)
```

```
## [1] "2004-01-01" "2004-02-01" "2004-03-01" "2004-04-01" "2004-05-01"
## [6] "2004-06-01"
```

Next, let's load packages that we will be using in the following sections.

```
# Load libraries
library(MASS)
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.3.3
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
# If you do not have these packages installed, please run the following line.
# Replace "PackageName" with the name of the package you'd like to install.
#install.packages("PackageName")
```

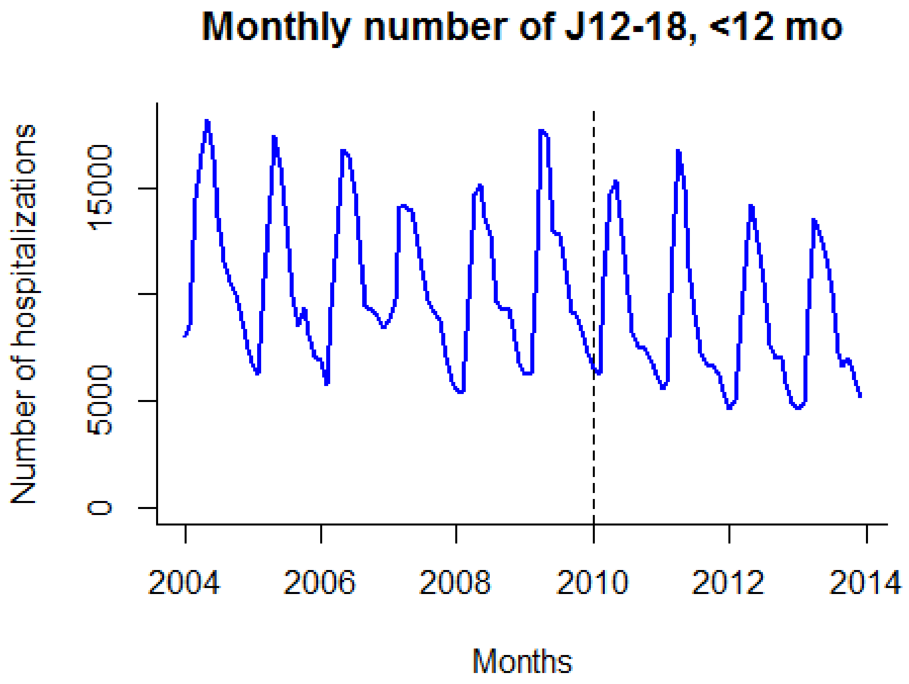
# Part 1. Visualize the Data

## Part 1-a. J12-18

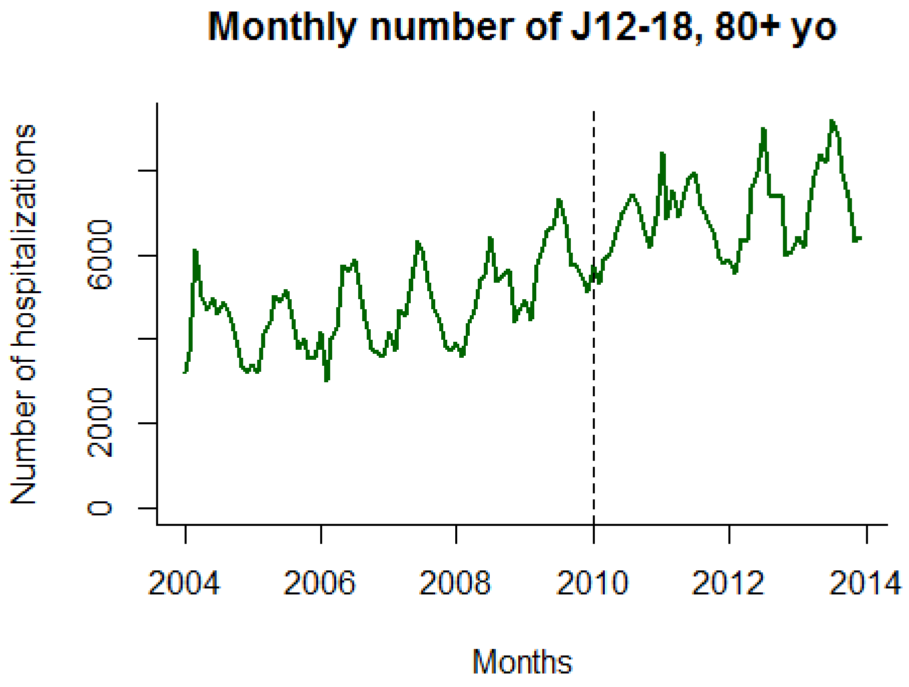
Make a plot for the time series for all-cause pneumonia hospitalizations (ICD10 code: J12-18) among children <12 months of age.

Sort the dataset by date, and make the same plots for <12 mo and 80+ yo.

```
# Sort the dataset by date
d <- d[order(d$date),]

# <12 mo
plot(J12_18 ~ date, data=d[d$age_group=="<1",],
     type="l", bty="l", col="blue", lwd=2,
     ylim=c(0,max(d$J12_18[d$age_group=="<1"])),
     xlab="Months", ylab="Number of hospitalizations",
     main="Monthly number of J12-18, <12 mo")
      abline(v=as.Date("2010-01-01"), lty=2)
```



```r
# 80+ yo
plot(J12_18 ~ date, data=d[d$age_group=="80+",],
     type="l", bty="l", col="darkgreen", lwd=2,
     ylim=c(0,max(d$J12_18[d$age_group=="80+"])),
     xlab="Months", ylab="Number of hospitalizations",
     main="Monthly number of J12-18, 80+ yo")
    abline(v=as.Date("2010-01-01"), lty=2)
```



What kind of trend do you see in J12-18 for each age group?

## Part 1-b. ACJ_NOJ

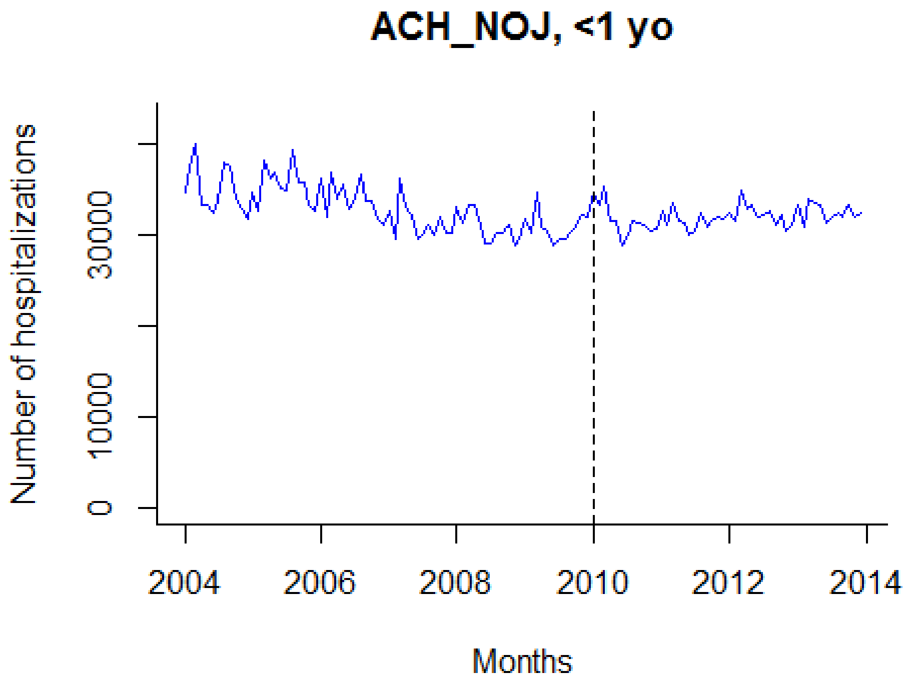
Plot the time series for non-respiratory hospitalizations (i.e., ACH_NOJ) for <12 mo and 80+ yo. This variable will be used as an offset for regression models.

First, to make the following analyses easier, let's subset the datasets into two age groups (<12 mo and 80+ yo).

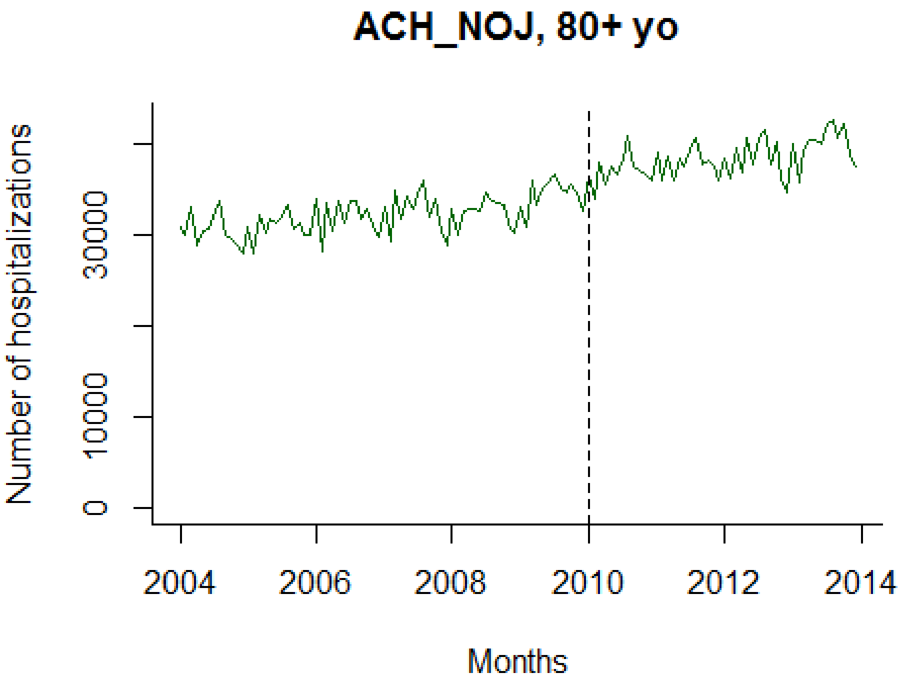
```
young <- d[d$age_group=="<1",]
old <- d[d$age_group=="80+",]
```

Now let's make plots for ACH_NOJ.

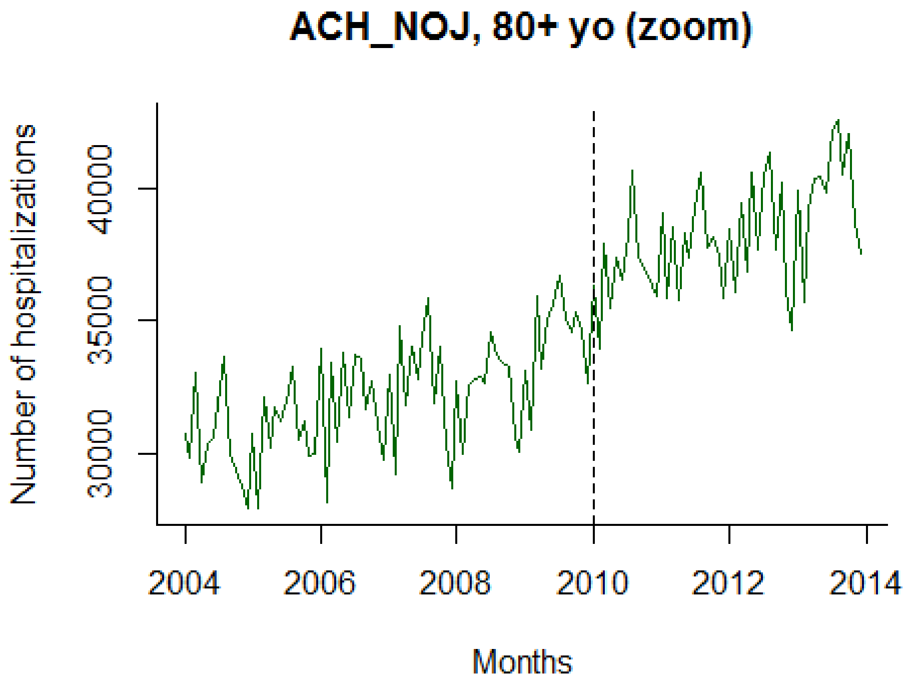
```
# <12 mo
plot(ACH_NOJ ~ date, data=young, bty="l", type="l",
     ylim=c(0,max(old$ACH_NOJ)),
     xlab="Months", ylab="Number of hospitalizations",
     col="blue", main="ACH_NOJ, <1 yo")
 abline(v=as.Date("2010-01-01"), lty=2)
```



```r
# 80+ yo
# Version 1 (y axis from zero to the max. number of hospitalizations)
plot(ACH_NOJ ~ date, data=old, bty="l", type="l",
     ylim=c(0,max(old$ACH_NOJ)),
     xlab="Months", ylab="Number of hospitalizations",
     col="darkgreen", main="ACH_NOJ, 80+ yo")
 abline(v=as.Date("2010-01-01"), lty=2)
```



```r
# Verson 2 (zoom in)
plot(ACH_NOJ ~ date, data=old, bty="l", type="l",
     xlab="Months", ylab="Number of hospitalizations",
     col="darkgreen", main="ACH_NOJ, 80+ yo (zoom)")
 abline(v=as.Date("2010-01-01"), lty=2)
```



What kind of trend do you see in ACH_NOJ for each age group?

## Part 2. Negative Binomial Regression

First, we will fit a regession just to the **pre-vaccne** data and will extrapolate the trend to the post-vaccine period to estimate the **counterfactual**.

Because the outcome is a **COUNT** variable, it is most appropriate to use a log-linked Poisson or negative binomial regression, rather than linear regression.

Due to the over-dispersion present in the data, we will fit a negative binonimal regression.

### Part 2-a. Set up

In order to fit a model just to the pre-vaccine period, set the outcome (J12-18) to missing (NA) for the post vaccine period.

NOTE: PCV10 was introduced on Jan 1, 2010 in Brazil.

```
# Create a new variable J12_18_pre which is NA (missing) in the post-vaccine
period.
# <12 mo
young$J12_18_pre <- young$J12_18
young$J12_18_pre[which(young$date>="2010-01-01")] <- NA
# 80+
old$J12_18_pre <- old$J12_18
old$J12_18_pre[which(old$date>="2010-01-01")] <- NA
```

```
# Check if it was created as we want.
#data.frame(young$date, young$J12_18, young$J12_18_pre)
#data.frame(old$date, old$J12_18, old$J12_18_pre)
```

Next, let's create on offset term for negative binomial regression using ACH_NOJ (in a log scale).

```
# Create an offset term in a log scale--this is the denominator for the
regression.
young$log_offset <- log(young$ACH_NOJ)
old$log_offset <- log(old$ACH_NOJ)
```

We will also create a time index variable to control for a long term linear trend.

```
# Create a time index variable (1, 2, 3, 4, ..., number of datapoints)
young$time <- 1:nrow(young)
old$time <- 1:nrow(old)
young$month<-as.factor(month(young$date))
old$month<-as.factor(month(old$date))
```

As the outcome J12-18 shows a clear seasonality, we will adjust for it in the regression model. We can do it in two ways: * Using monthly dummy variables (We will do this here) * Using harmonic terms (sine, cosine) ### Part 2-b. Fit a negative binomial model

Fit negative binomial models to the prevaccine data.

```
NB_yng_s1 <- glm.nb(J12_18_pre ~ time + month         + offset(log_offset),
data=young)
NB_old_s1 <- glm.nb(J12_18_pre ~ time + month         + offset(log_offset),
data=old)
```

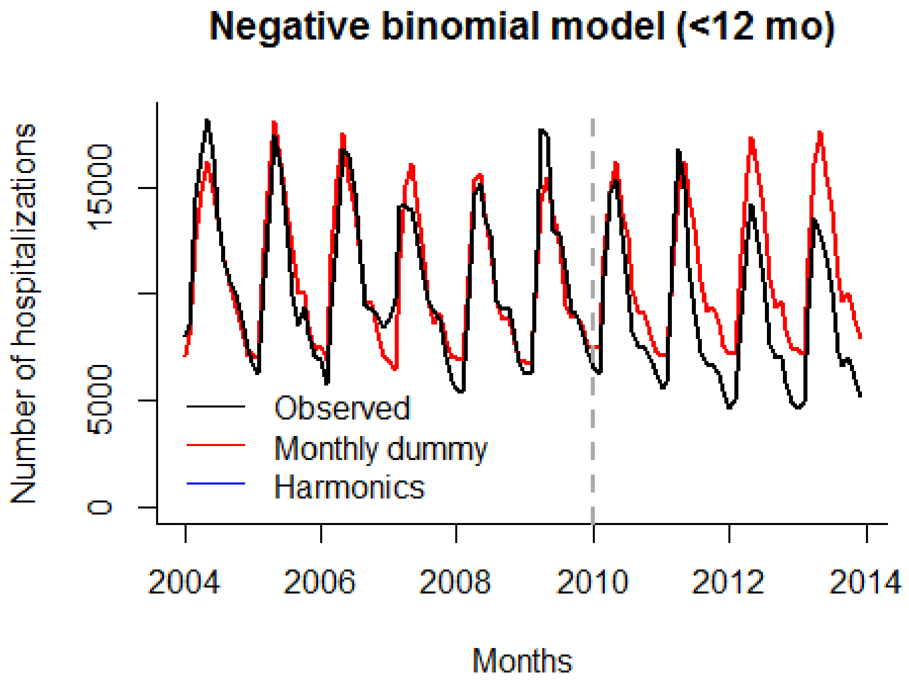
Exrapolate the trend to the post-vaccine period and generate the counterfactual for J12-18.

```
Pred_NB_yng_s1 <- predict(NB_yng_s1, newdata=young, type="response",
se.fit=T)
#Pred_NB_yng_s2 <- predict(NB_yng_s2, newdata=young, type="response",
se.fit=T)
Pred_NB_old_s1 <- predict(NB_old_s1, newdata=old,   type="response",
se.fit=T)
#Pred_NB_old_s2 <- predict(NB_old_s2, newdata=old,   type="response",
se.fit=T)
```

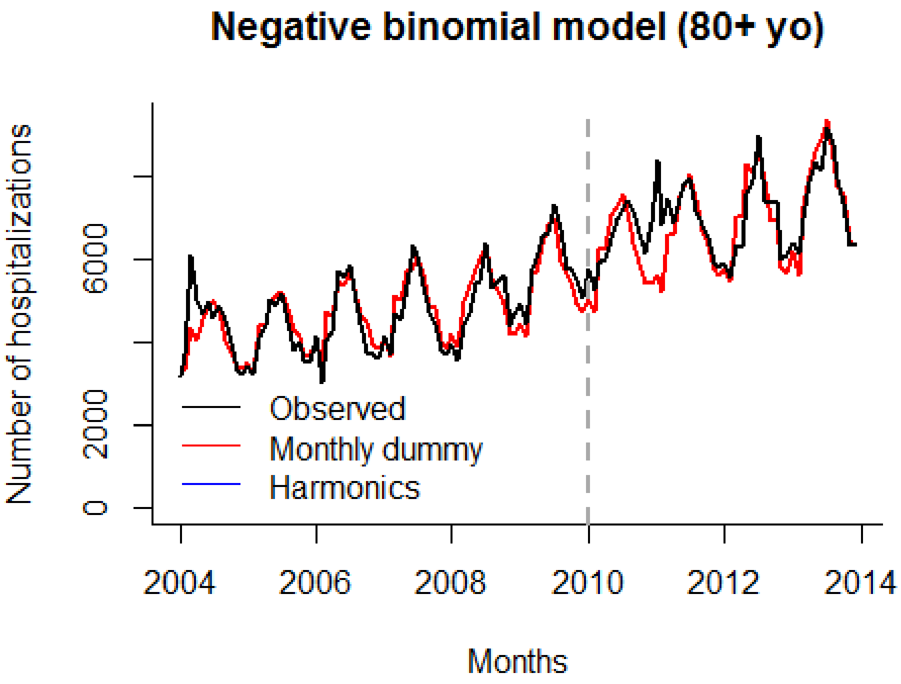
Plot time series for observed J12-18 vs. counterfactual J12-18.

```
# <12 mo
plot(Pred_NB_yng_s1$fit ~ young$date,
     type="l",col="red", bty="l", lwd=2,
     ylim=c(0,max(c(young$J12_18, Pred_NB_yng_s1$fit))),
     ylab="Number of hospitalizations", xlab="Months",
     main="Negative binomial model (<12 mo)")
lines(J12_18 ~ date, data=young, col="black",lwd=2)
```

```
abline(v=as.Date("2010-01-01"),col="darkgrey",lty=2, lwd=2)
legend(x="bottomleft",legend=c("Observed","Monthly dummy","Harmonics"),
       col=c("black","red","blue"),lty=c(1,1,1),bty="n")
```



```
# 80+ mo
plot(Pred_NB_old_s1$fit ~ old$date,
     type="l",col="red", bty="l", lwd=2,
     ylim=c(0,max(c(old$J12_18, Pred_NB_old_s1$fit))),
     ylab="Number of hospitalizations", xlab="Months",
     main="Negative binomial model (80+ yo)")
lines(J12_18 ~ date, data=old, col="black",lwd=2)
abline(v=as.Date("2010-01-01"),col="darkgrey",lty=2, lwd=2)
legend(x="bottomleft",legend=c("Observed","Monthly dummy","Harmonics"),
       col=c("black","red","blue"),lty=c(1,1,1),bty="n")
```



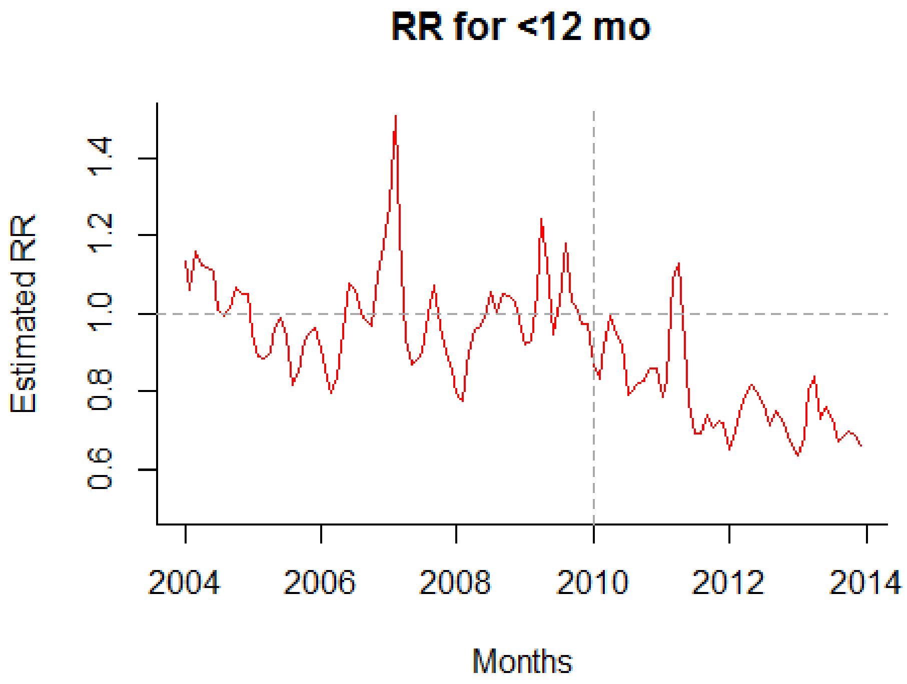
## Part 2-c. Rate ratios (RRs)

Calculate the rate ratios (RRs)

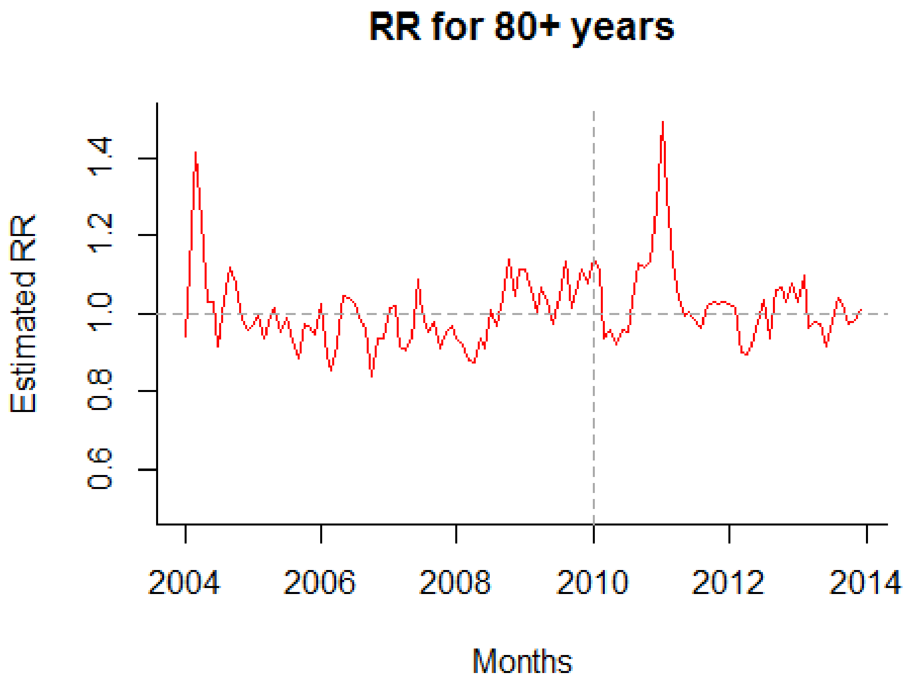
```
RR_NB_yng <- young$J12_18/Pred_NB_yng_s1$fit
RR_NB_old <- old$J12_18/Pred_NB_old_s1$fit
```

Plot RRs by time.

```
# <12 mo
plot(RR_NB_yng ~ young$date, type="l", bty="l", col="red",
     main="RR for <12 mo", xlab="Months", ylab="Estimated RR",
ylim=c(0.5,1.5))
abline(h=1,col="darkgrey",lty=2)
abline(v=as.Date("2010-01-01"),col="darkgrey",lty=2)
```



```
# 80+ yo
plot(RR_NB_old ~ old$date, type="l", bty="l", col="red",
     main="RR for 80+ years", xlab="Months", ylab="Estimated RR",
ylim=c(0.5,1.5))
abline(h=1,col="darkgrey",lty=2)
abline(v=as.Date("2010-01-01"),col="darkgrey",lty=2)
```



## Part 2-e. Leave-one-season-out analysis

As a sensitivity analysis, we can fit a series of negative binomial models by excluding one season at a time.

For example, the 1st model will be fit to the pre-vaccine data excluding the first year of the pre-vaccine period; the 2nd model will be fit to the pre-vaccine data excluding the second year...

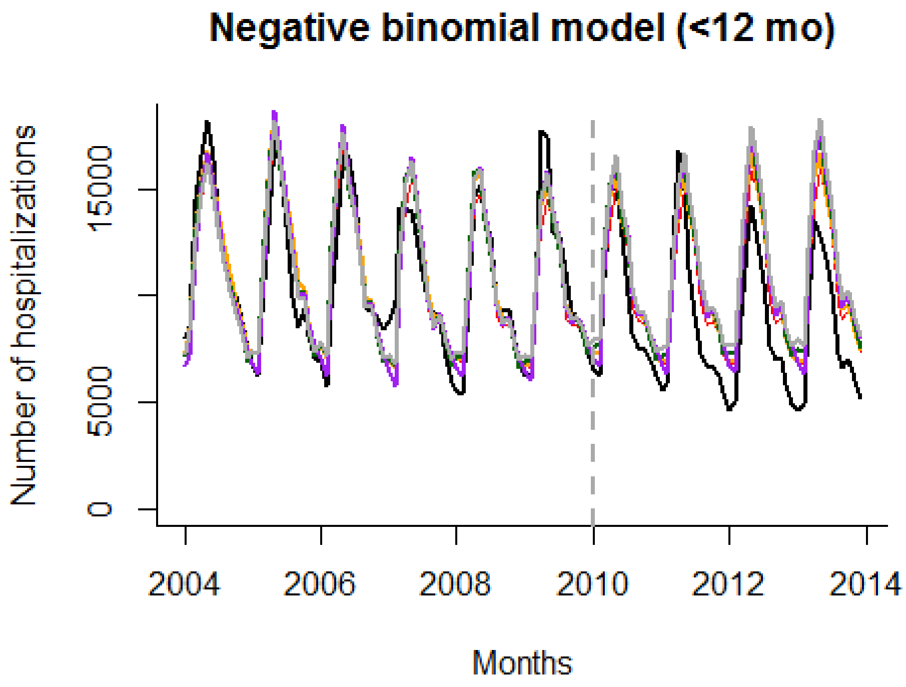
There are 6 years of pre-vaccine data, so we will fit 6 models.

Let's start with the young age group.

```
# First, let's create an empty matrix to store results.
lvso_yng <- matrix(NA, nrow=nrow(young), ncol=6)
for (i in 1:6) {

  # 1. Create J12_18_pre as before
  young$J12_18_pre <- young$J12_18
  young$J12_18_pre[which(young$date>="2010-01-01")] <- NA

  # 2. Exclude one season from the pre-vaccine period
  k <- (12*(i-1)+1):(12*(i-1)+12)
  young$J12_18_pre[k] <- NA

  # 3. Fit a negative binomial model
  NB_yng_lvso <- glm.nb(J12_18_pre ~ time+month+offset(log_offset),
data=young)

  # 4. Extrapolate a trend to the post-vaccine period
  Pred_NB_yng_lvso <- predict(NB_yng_lvso, newdata=young, type="response",
se.fit=T)

  # 5. Save a result in a matrix
  lvso_yng[,i] <- Pred_NB_yng_lvso$fit
}
```

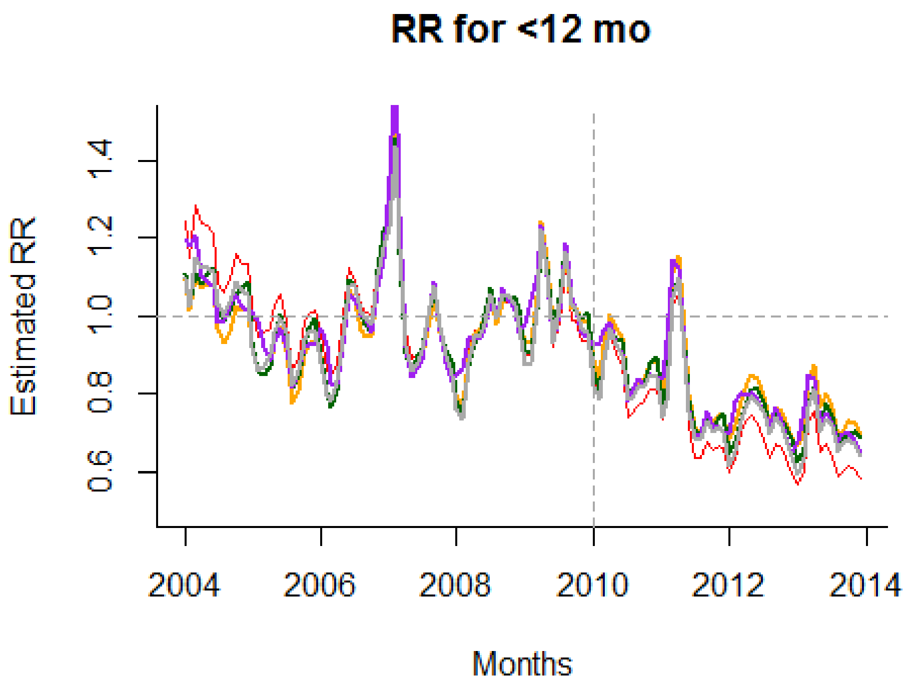
Plot observed vs. counterfactual.

```
plot(J12_18 ~ date, data = young,
     type="l",col="black", bty="l", lwd=2,
     ylim=c(0,max(c(young$J12_18))),
     ylab="Number of hospitalizations", xlab="Months",
     main="Negative binomial model (<12 mo)")
lines(lvso_yng[,i] ~ young$date, data=young, col="red")
col <- c("blue","orange","darkgreen","purple","darkgrey")
for (i in 2:6) {
  lines(lvso_yng[,i] ~ date, data=young, col=col[i],lwd=2)
}
abline(v=as.Date("2010-01-01"),col="darkgrey",lty=2, lwd=2)
```



Calculate and plot RRs.

```
RR_lvso_yng <- young$J12_18/lvso_yng
plot(RR_lvso_yng[,1] ~ young$date, type="l", bty="l", col="red",
     main="RR for <12 mo", xlab="Months", ylab="Estimated RR",
ylim=c(0.5,1.5))
for (i in 2:6) {
  lines(RR_lvso_yng[,i] ~ date, data=young, col=col[i],lwd=2)
}
abline(h=1,col="darkgrey",lty=2)
abline(v=as.Date("2010-01-01"),col="darkgrey",lty=2)
```



## Part 3. Interrupted Time Series Analysis

In this section, we compare a simple univariate linear regression with interupted time series regressions where we test whether the slope of the line changes after vaccine introduction.

### Part 3-a. Set up

First, let's create the following dummy variables for the post-vaccine period. * period1: 1 if 1-12 months after PCV10 introduction * period2: 1 if >12 months after PCV10 introduction

```
# <12 mo
young$period1 <- 0
young$period2 <- 0
young$period1[young$date>="2010-01-01" & young$date<"2011-01-01"] <- 1
young$period2[young$date>="2011-01-01"] <- 1
```

```
# 80+ yo
old$period1 <- 0
old$period2 <- 0
old$period1[old$date>="2010-01-01" & old$date<"2011-01-01"] <- 1
old$period2[old$date>="2011-01-01"] <- 1
```

## Part 3-b. Fit 3 models

Fit an interrupted time series model as follows.

```
ITS_yng <- glm.nb(J12_18 ~ month + time*period1 + time*period2, data=young)
ITS_old <- glm.nb(J12_18 ~ month + time*period1 + time*period2, data=old)
```

NOTE: This model includes time, period1, and period2 althouth these terms are not explicitly written in the code above.

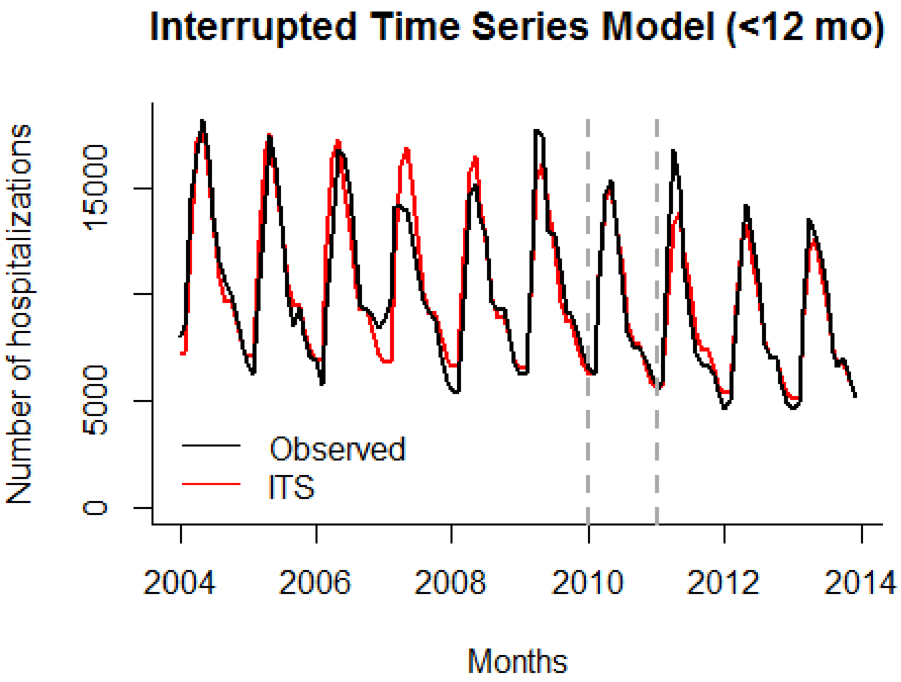
## Part 3-c. Plot fitted values

Calculate fitted values.

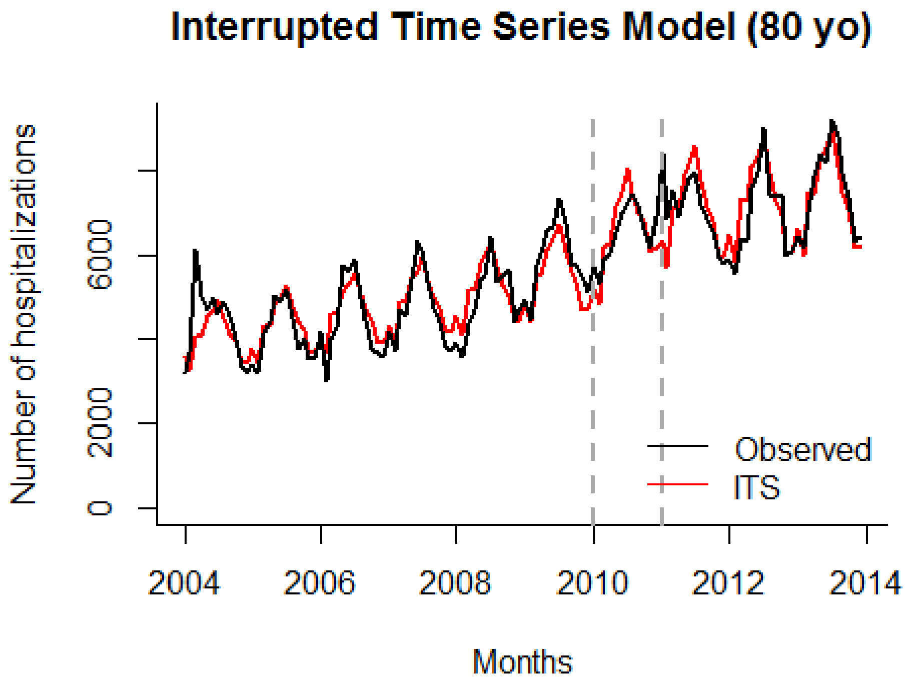
```
Pred_ITS_yng <- predict(ITS_yng, newdata=young, type="response", se.fit=T)
Pred_ITS_old <- predict(ITS_old, newdata=old,   type="response", se.fit=T)
```

Make plots for the observed vs. fitted.

```
# <12 mo
plot(Pred_ITS_yng$fit ~ young$date,
     type="l",col="red", bty="l", lwd=2,
     ylim=c(0,max(young$J12_18)),
     ylab="Number of hospitalizations", xlab="Months",
     main="Interrupted Time Series Model (<12 mo)")
lines(J12_18 ~ date, data=young, col="black",lwd=2)
abline(v=as.Date("2010-01-01"),col="darkgrey",lty=2, lwd=2)
abline(v=as.Date("2011-01-01"),col="darkgrey",lty=2, lwd=2)
legend("bottomleft",
legend=c("Observed","ITS"),col=c("black","red"),lty=c(1,1),bty="n")
```



```r
# 80+ yo
plot(Pred_ITS_old$fit ~ young$date,
     type="l",col="red", bty="l", lwd=2,
     ylim=c(0,max(old$J12_18)),
     ylab="Number of hospitalizations", xlab="Months",
     main="Interrupted Time Series Model (80 yo)")
lines(J12_18 ~ date, data=old, col="black",lwd=2)
abline(v=as.Date("2010-01-01"),col="darkgrey",lty=2, lwd=2)
abline(v=as.Date("2011-01-01"),col="darkgrey",lty=2, lwd=2)
legend("bottomright",
legend=c("Observed","ITS"),col=c("black","red"),lty=c(1,1),bty="n")
```



## Part 3-d. Generate counterfactual and estimate the impact of PCV10

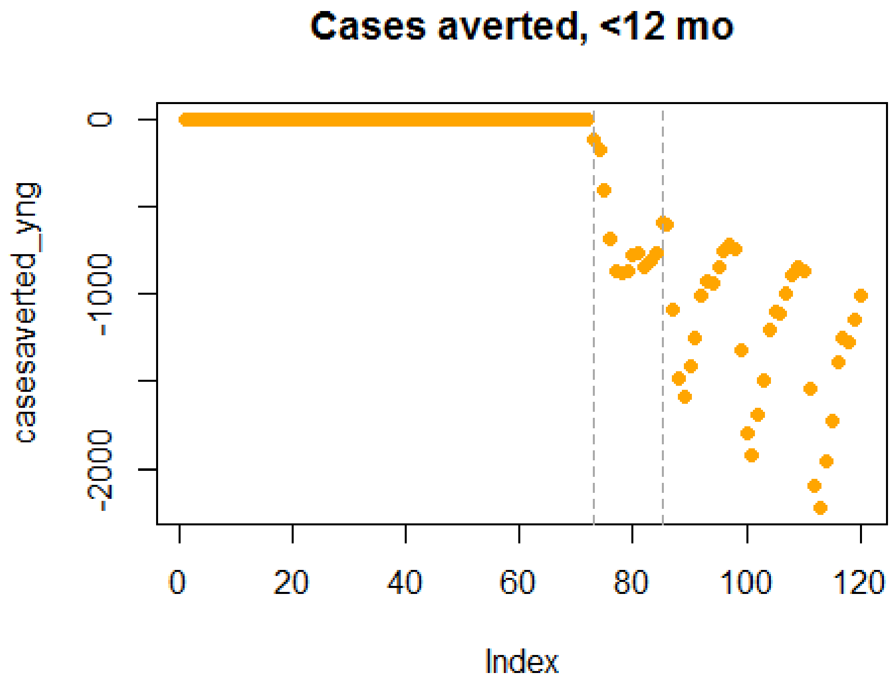
Calculate the counterfactual which is the number of cases expected without PCV10. In this case, that's the following part of the model:

$$\beta_0 + \beta_1 * time + \beta_2 * sin12 + \beta_3 * cos12$$

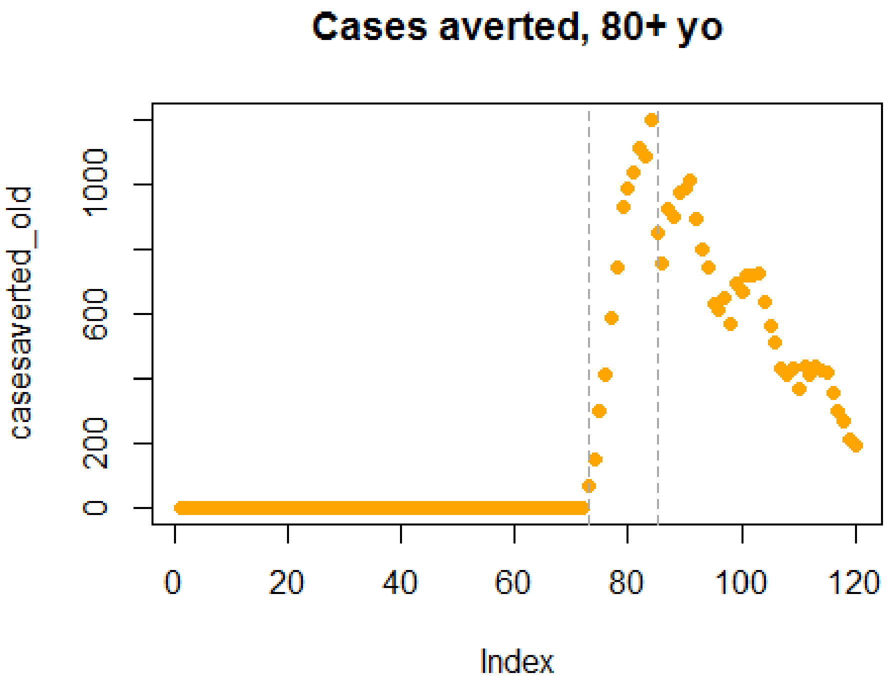
```
# < 12mo
#cf_yng <-exp(ITS_yng$coef[1] + ITS_yng$coef[2]*young$sin12 +
ITS_yng$coef[3]*young$cos12 + ITS_yng$coef[4]*young$time)
cf_yng<- Pred_ITS_yng$fit / exp(young$period1*ITS_yng$coef['period1'] +
young$period2*ITS_yng$coef['period2']

+young$period1*young$time*ITS_yng$coef['time:period1']

+young$period2*young$time*ITS_yng$coef['time:period2'] )
# 80+ yo
cf_old<- Pred_ITS_old$fit / exp(old$period1*ITS_old$coef['period1']
+old$period2*ITS_old$coef['period2']

+old$period1*old$time*ITS_old$coef['time:period1']

+old$period2*old$time*ITS_old$coef['time:period2'] )
```

Calculate and plot the number of cases averted.

```r
# First, let's reformat ITS_###3$fit as follows:
str(Pred_ITS_yng$fit) # It is a "named number", so let's unmane them

##  Named num [1:120] 7226 7276 12810 17105 17945 ...
##  - attr(*, "names")= chr [1:120] "121" "122" "123" "124" ...

Pred_ITS_yng <- unname(Pred_ITS_yng$fit)
Pred_ITS_old <- unname(Pred_ITS_old$fit)

# Calculate the number of cases we averted using our intervention
casesaverted_yng <- Pred_ITS_yng - cf_yng
casesaverted_old <- Pred_ITS_old - cf_old

# Plot
plot(casesaverted_yng, col="orange", main="Cases averted, <12 mo", pch=16)
abline(v=73, col="darkgrey", lty=2)
abline(v=85, col="darkgrey", lty=2)
```

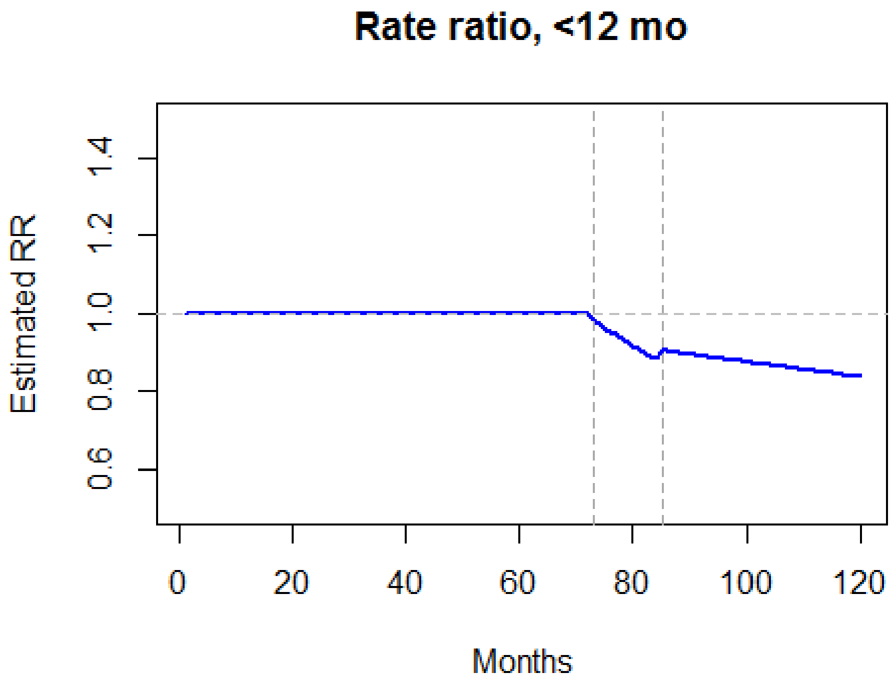


```r
plot(casesaverted_old, col="orange", main="Cases averted, 80+ yo", pch=16)
abline(v=73, col="darkgrey", lty=2)
abline(v=85, col="darkgrey", lty=2)
```

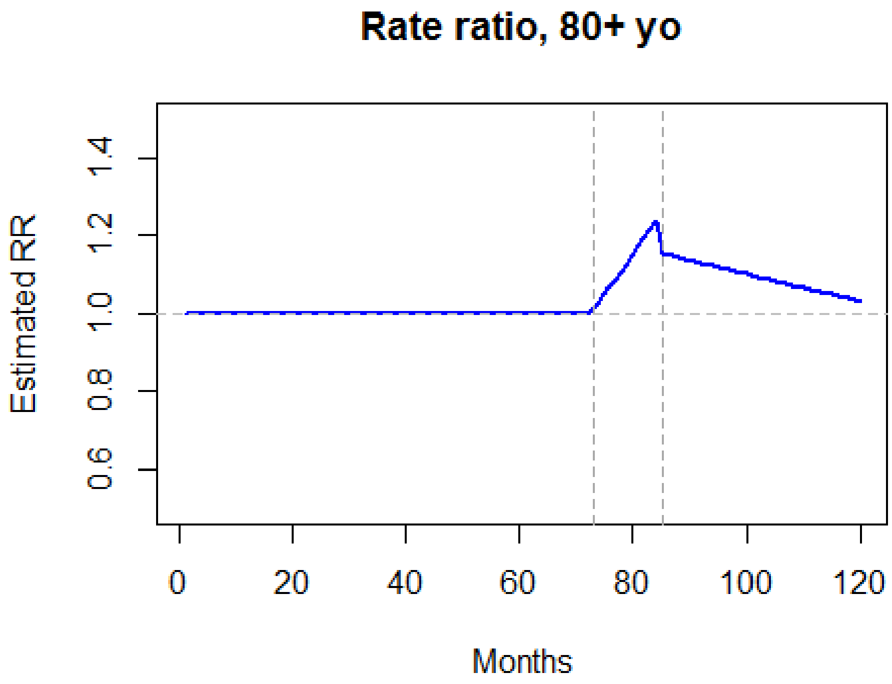


What about the change in rate?

```
# <12 mo
RR_yng <- Pred_ITS_yng/cf_yng
plot(RR_yng, type="l", col="blue", lwd=2, main="Rate ratio, <12 mo",
     xlab="Months",ylab="Estimated RR", ylim=c(0.5, 1.5))
abline(v=73, col="darkgrey", lty=2)
abline(v=85, col="darkgrey", lty=2)
abline(h=1, col='gray', lty=2)
```



```
# 80+
RR_old <- Pred_ITS_old/cf_old
plot(RR_old, type="l", col="blue", lwd=2, main="Rate ratio, 80+ yo",
     xlab="Months",ylab="Estimated RR", ylim=c(0.5, 1.5))
abline(v=73, col="darkgrey", lty=2)
abline(v=85, col="darkgrey", lty=2)
abline(h=1, col='gray', lty=2)
```



Thank you for your participation! Please feel free to contact us anytime if you have any questions! Daniel M. Weinberger (daniel.weinberger@yale.edu) and Kayoko Shioda (kayoko.shioda@yale.edu)