Exercise sheet 5

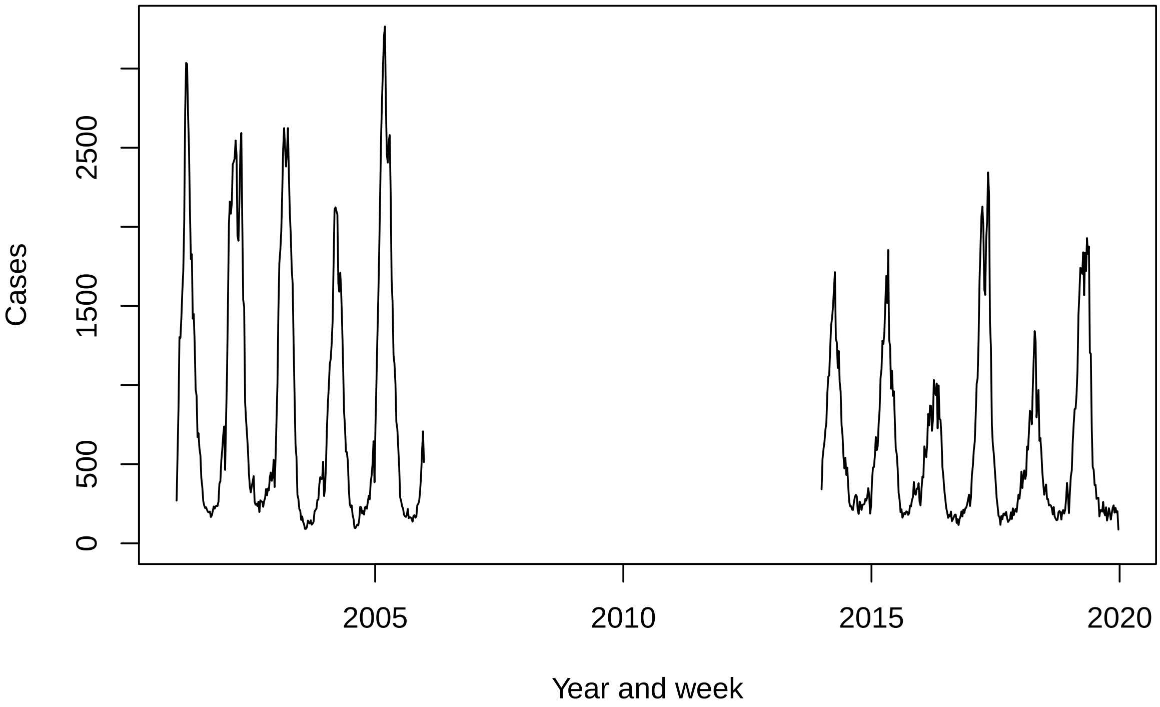
Prepare the below such that you are able to discuss it on Wednesday 31st March

Exercise 1 Interrupted time series analysis

A simple method of determining whether an intervention is useful is to examine cases observed before and after the introduction of the intervention and see whether the two periods are different, specifically whether there is a drop in infectious disease cases after introduction of the measure. This kind of approach is known as interrupted time series analysis or change-point analysis. In Germany, vaccines for rotavirus were made available in the mid 2000s and included in the national immunisation schedule in 2013. We want to know what effect a rotavirus infection vaccine might have had on case numbers. For this, we analyse SurvStat@RKI data which provide us with the reported number of laboratory confirmed cases for each week since rotavirus became a notifiable disease in Germany in 2001.

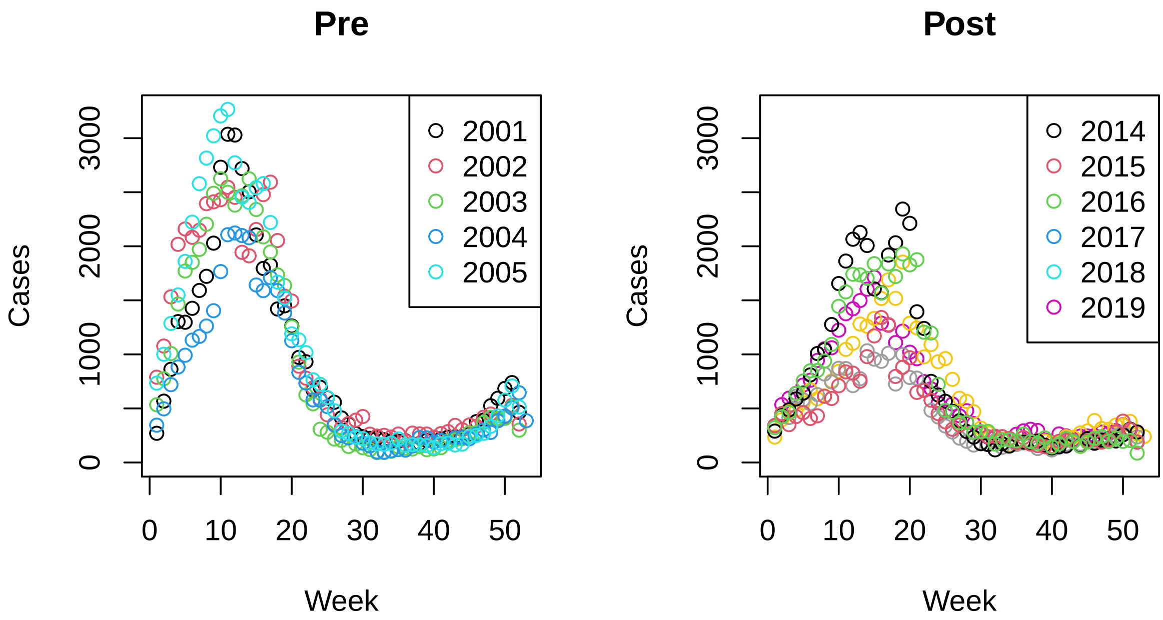
a) Load the data set `rota` from OLAT and plot the data for the period before 2006 (pre) and after 2013 (post). Use `ISOweek2date` from the **ISOweek** package to adjust the case notification dates. What does the trend look like for the two periods?
Hint: You may wish to remove 2020 and 2021 data to avoid COVID-19 effects (social distancing and stay at home measures) and incomplete year of reporting, respectively.
Solution:

```
load("data/rota.rdata")
library(lubridate)

##
## Attaching package:  'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

rota$Notification <- ISOweek::ISOweek2date(gsub("w", "W",
                                    paste0(rota$Notification, "-1")))
covid <- rota[rota$Year >= 2020, ]
rota <- rota[isoyear(rota$Notification) < 2020, ]
rota[!(isoyear(rota$Notification) < 2006 |
               isoyear(rota$Notification) > 2013), ]$Cases <- NA
rota$Vax <- isoyear(rota$Notification) > 2013

with(rota, plot(Notification, Cases, type = "l",
                ylim = c(0, max(rota$Cases, na.rm = TRUE)),
                xlim = range(rota$Notification, na.rm = TRUE),
                xlab = "Year and week"))
```



There may be slightly fewer cases in the post period but it is hard to determine from visual inspection alone. We see seasonality in the data with most cases expected around this time of year/in the spring. We see fewer cases in the post period (the peak is smaller) and they seem to be occuring slightly later in the year.

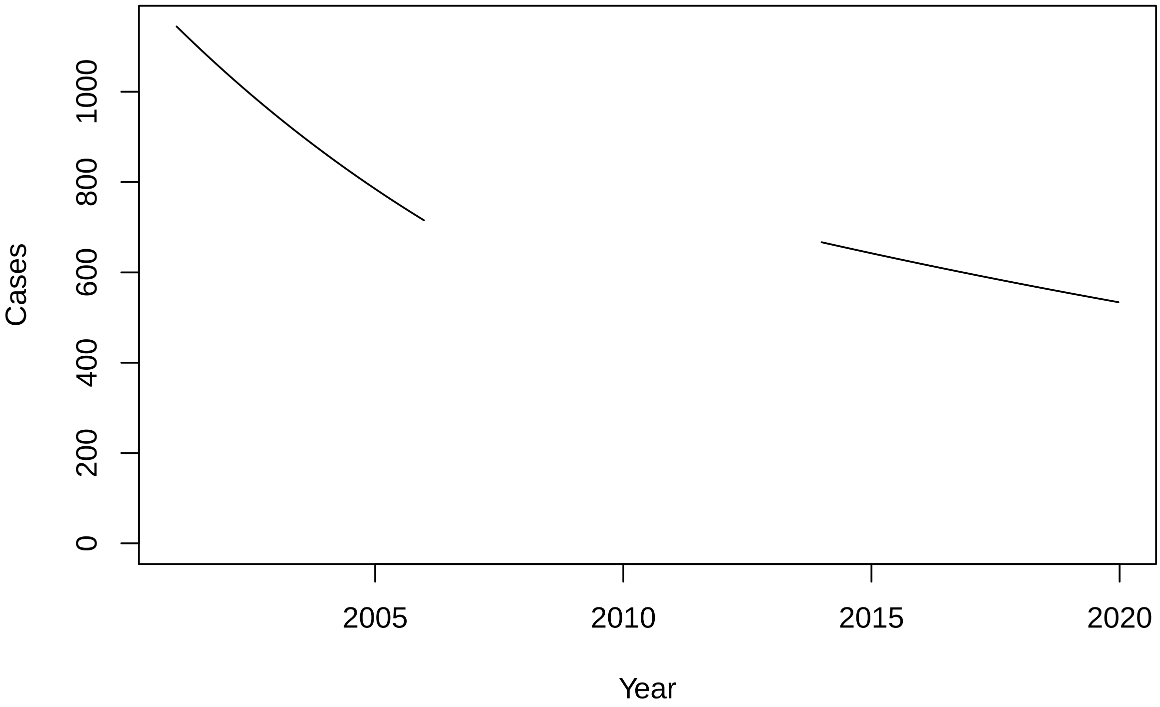
```
par(mfrow = c(1, 2))
with(rota[isoyear(rota$Notification) < 2006, ],
     plot(Week, Cases, col = Year, main = "Pre",
          ylim = c(0, max(rota$Cases, na.rm = TRUE))))
legend("topright", legend = unique(rota[isoyear(rota$Notification) < 2006, ]$Year),
       col = 1 : length(unique(rota[isoyear(rota$Notification) < 2006, ]$Year)),
       pch = 1)
with(rota[isoyear(rota$Notification) > 2013, ],
     plot(Week, Cases, col = Year, main = "Post",
          ylim = c(0, max(rota$Cases, na.rm = TRUE))))
legend("topright", legend = unique(rota[isoyear(rota$Notification) > 2013, ]$Year),
       col = 1 : length(unique(rota[isoyear(rota$Notification) > 2013, ]$Year)),
       pch = 1)
```



b) Fit a simple GLM model with Poisson link to the observed number of cases both periods and comment on your findings.

Solution:

```
m <- glm(Cases ~ Notification * Vax, data = rota, family = poisson,
         na.action = na.exclude)

idx <- which(is.na(rota$Cases))
plot(rota$Notification,
     # pad with NAs to obtain same length
     c(m$fitted.values[1 : idx[1] - 1], rep(NA, length(idx)),
       m$fitted.values[idx[1] : length(m$fitted.values)]),
     xlim = range(rota$Notification),
     type = "l", ylim = c(0, max(m$fitted.values)), ylab = "Cases",
     xlab = "Year")
```



The slope seems to be different in the pre and post intervention periods. We determine this both visually (the curves look different) but also based on estimated model parameters. The test below also shows a difference between the two models (with and without interaction with `Vax`, the variable determining when the vaccination has been introduced):

```
summary(m)

##
## Call:
## glm(formula = Cases ~ Notification * Vax, family = poisson, data = rota,
##     na.action = na.exclude)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -34.57  -20.22  -12.04   12.17   66.66
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            9.966e+00  4.754e-02  209.66   <2e-16 ***
## Notification         -2.582e-04  3.905e-06  -66.12   <2e-16 ***
## VaxTRUE              -1.830e+00  7.867e-02  -23.27   <2e-16 ***
## Notification:VaxTRUE  1.566e-04  5.352e-06   29.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 329670  on 573  degrees of freedom
## Residual deviance: 305515  on 570  degrees of freedom
##   (417 observations deleted due to missingness)
## AIC: 310132
##
## Number of Fisher Scoring iterations: 5

m2 <- glm(Cases ~ Notification, data = rota, family = poisson)
anova(m, m2, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: Cases ~ Notification * Vax
## Model 2: Cases ~ Notification
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       570     305515
## 2       572     307427 -2  -1911.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c) What might be a drawback of such an approach?
Solution: We have to be somewhat certain that the intervention(s) of interest are not confounded with other things occuring simultaneously to be able to conclude the difference in slopes is due to the intervention. For example, a lot of measures were used in malaria elimination efforts in the early 20th century so it would be difficult to disentangle the effects of any of them:

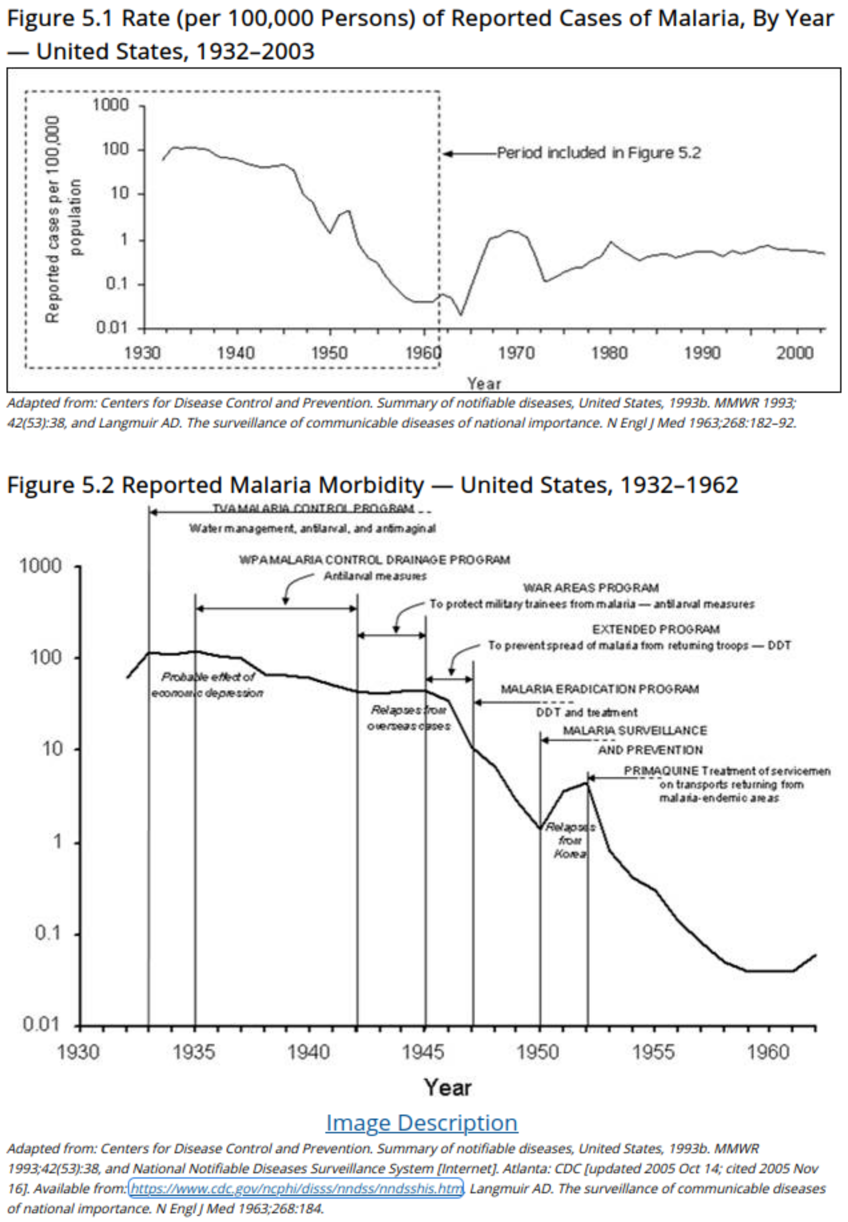


Figure 5.1 Rate (per 100,000 Persons) of Reported Cases of Malaria, By Year — United States, 1932–2003

Adapted from: Centers for Disease Control and Prevention. Summary of notifiable diseases, United States, 1993b. MMWR 1993; 42(53):38, and Langmuir AD. The surveillance of communicable diseases of national importance. N Engl J Med 1963;268:182–92.

Figure 5.2 Reported Malaria Morbidity — United States, 1932–1962

TVA MALARIA CONTROL PROGRAM
Water management, antilarval, and antimaginal
WPA MALARIA CONTROL DRAINAGE PROGRAM
Antilarval measures
WAR AREAS PROGRAM
To protect military trainees from malaria — antilarval measures
EXTENDED PROGRAM
To prevent spread of malaria from returning troops — DDT
MALARIA ERADICATION PROGRAM
DDT and treatment
MALARIA SURVEILLANCE
AND PREVENTION
PRIMAQUINE Treatment of servicemen on transports returning from malaria-endemic areas
Probable effect of economic depression
Relapses from overseas cases
Relapses from Korea
1000
100
10
1
0.1
0.01
1930
1935
1940
1945
1950
1955
1960
Year

Image Description

Adapted from: Centers for Disease Control and Prevention. Summary of notifiable diseases, United States, 1993b. MMWR 1993;42(53):38, and National Notifiable Diseases Surveillance System [Internet]. Atlanta: CDC [updated 2005 Oct 14; cited 2005 Nov 16]. Available from: https://www.cdc.gov/ncphi/disss/nndss/nndsshis.htm, Langmuir AD. The surveillance of communicable diseases of national importance. N Engl J Med 1963;268:184.

Image source: Principles of Epidemiology in Public Health Practice (USCDC) Lesson 5 section 5 (`https://www.cdc.gov/csels/dsepd/ss1978/lesson5/section5.html`)
This is why it was suggested you remove cases after 2019, as they seem to be effected by ongoing public health interventions seeking to control COVID-19 (and not anything else specific to rotavirus as far as we are aware)

```
range(covid$Cases)

## [1]  37 329

range(rota$Cases, na.rm = TRUE)

## [1]   87 3266
```

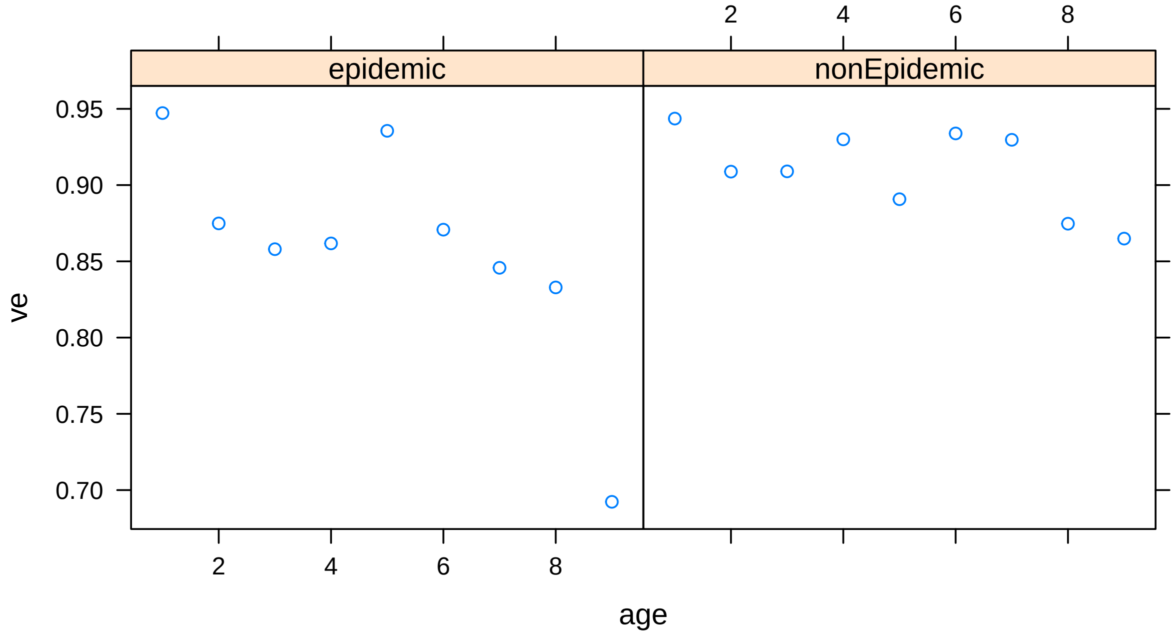
Additionally, there is an increase in cases before the interruption which is due to artefacts that are not immunisation-related (see https://doi.org/10.1093/biostatistics/kxy057 for details). There might also have been changes in population following increased migration from 2015. A change in the birthrate would have an effect on cases, e.g. if birthrates go down, the number of children will decrease, and so also will cases.

Exercise 2 Pertussis

The file `whoopingCough` available on OLAT contains data on pertussis in the UK collected over two periods, one just prior to the start ("non-epidemic") and one at the first peak of the pertussis epidemic ("epidemic") of 1989–1990. The following information is available: reporting period (`period`), 1-year age groups (`age`), total number of notified cases (`cases`), number of vaccinated cases (`vaccinated`), vaccination coverage figures for the relevant birth cohorts as a proxy for the proportions of vaccinated 1–9 year olds in the population (`coverage`).

a) Compute the vaccine effectiveness for each age-group and period. Create a plot for each period where vaccine effectiveness is shown as a function of age.
Solution:

```
whoopingCough <- read.table("data/whoopingCough.txt", header = TRUE, sep = ",")
whoopingCough <- within(whoopingCough, {
        # proportion of the population vaccinated
        ppv <- coverage / 100
        # proportion of the cases vaccinated
        pcv <- vaccinated / cases
        # vaccine effectiveness
        ve <- 1 - pcv / (1 - pcv) * (1 - ppv) / ppv
        # offset for screening method
        o <- qlogis(ppv)
})
library(lattice)
(xyplot(ve ~ age | period, data = whoopingCough))
```



b) Fit a suitable binomial GLM to the data with a two-level factor for the study period and a nine-level factor for age as main effects. Use Pearson's χ^2 statistic to estimate overdispersion ϕ for this model.
Solution: We consider additive effects in this model

```
m <- glm(cbind(vaccinated, cases - vaccinated) ~ offset(o) + period +
                 as.factor(age), family = binomial, data = whoopingCough)
summary(m)

##
## Call:
## glm(formula = cbind(vaccinated, cases - vaccinated) ~ offset(o) +
##     period + as.factor(age), family = binomial, data = whoopingCough)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.10969  -0.76295  -0.00009   0.64013   2.04151
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.7628     0.1619 -17.066  < 2e-16 ***
## periodnonEpidemic   -0.2933     0.1075  -2.729 0.006360 **
## as.factor(age)2      0.6723     0.1995   3.370 0.000753 ***
## as.factor(age)3      0.7461     0.2009   3.714 0.000204 ***
## as.factor(age)4      0.6118     0.2066   2.961 0.003069 **
## as.factor(age)5      0.4642     0.2069   2.243 0.024865 *
## as.factor(age)6      0.5568     0.2339   2.380 0.017301 *
## as.factor(age)7      0.7179     0.2720   2.639 0.008314 **
## as.factor(age)8      0.9765     0.3007   3.248 0.001163 **
## as.factor(age)9      1.3443     0.3170   4.241 2.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 50.546  on 17  degrees of freedom
## Residual deviance: 15.308  on  8  degrees of freedom
## AIC: 118.85
##
## Number of Fisher Scoring iterations: 4
```

The estimate of the overdispersion $\hat{\phi}$ is

```
(phi <- sum(residuals(m, type = "pearson") ^ 2) / m$df.residual)

## [1] 1.904082
```

c) Are the effects of `period` and the factor `age` as a whole significant at the $\alpha = 0.05$ level after adjusting for overdispersion?
Solution: Age is no longer significant (and period is borderline)

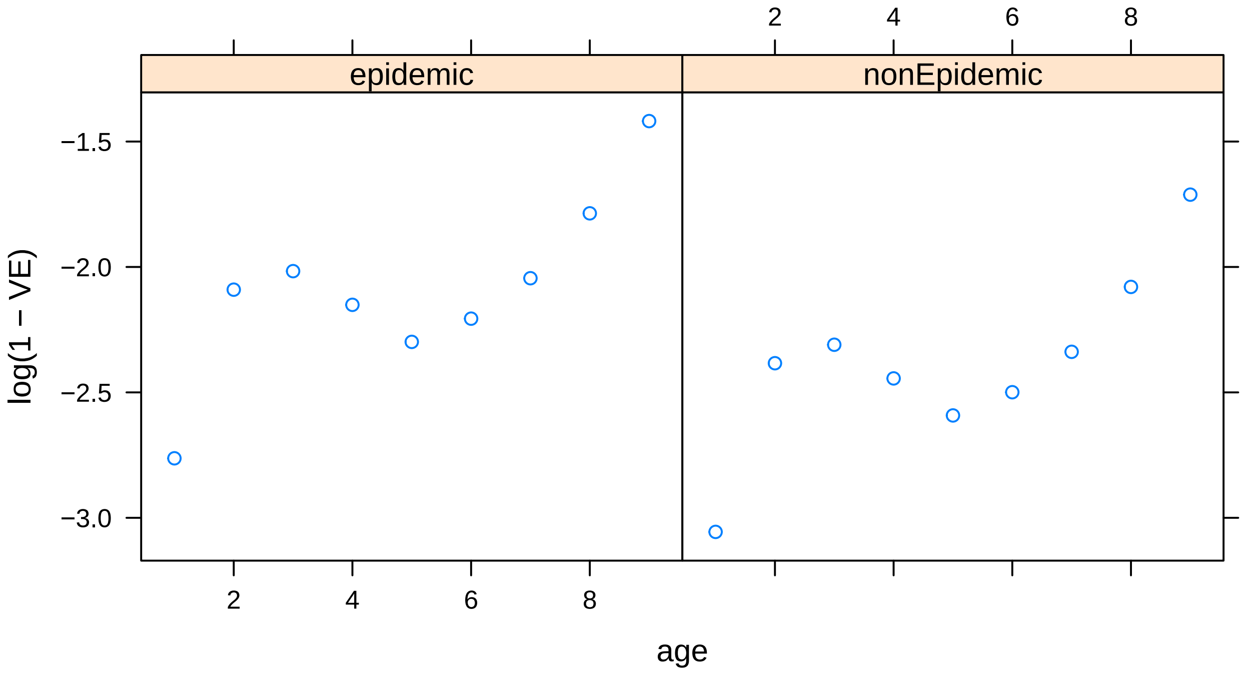
```
anova(m, test = "Chisq", dispersion = phi)

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
```

```
## Response: cbind(vaccinated, cases - vaccinated)
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                               17     50.546
## period           1   7.5367        16     43.009  0.04664 *
## as.factor(age)   8  27.7016         8     15.308  0.06854 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) Compute the estimated period- and age-specific vaccine effectiveness from your GLM model. Plot the resulting $\log(1-\text{VE})$ against age for each study period. What do you see?

Solution:

```
VE <- 1 - exp(predict(m) - whoopingCough$o)
(xyplot(log(1 - VE) ~ age | period, data = whoopingCough))
```



The patterns are more similar than in question a).