Infectious_diseases

VC

5/13/2019

Infectious diseases are major concern in densely populated conditions. Understandig trends of infectious diseases prevented by vaccine can give a good insight on what and how the diseases spread and what conditions diseases depend on. Here I have gathered infectious disease data of California from CDC, and demography data from 2010 census (https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF)

```
##needed packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0     v readr   1.3.1
## v tibble  2.0.1     v purrr   0.3.0
## v tidyr   0.8.2     v stringr 1.4.0
## v ggplot2 3.1.0     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(ggpubr)
```

```
## Loading required package: magrittr

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
###
census<- read.csv("DEC_10_SF1_GCTPH1.ST05_with_ann.csv",header=T)
head(census)
```

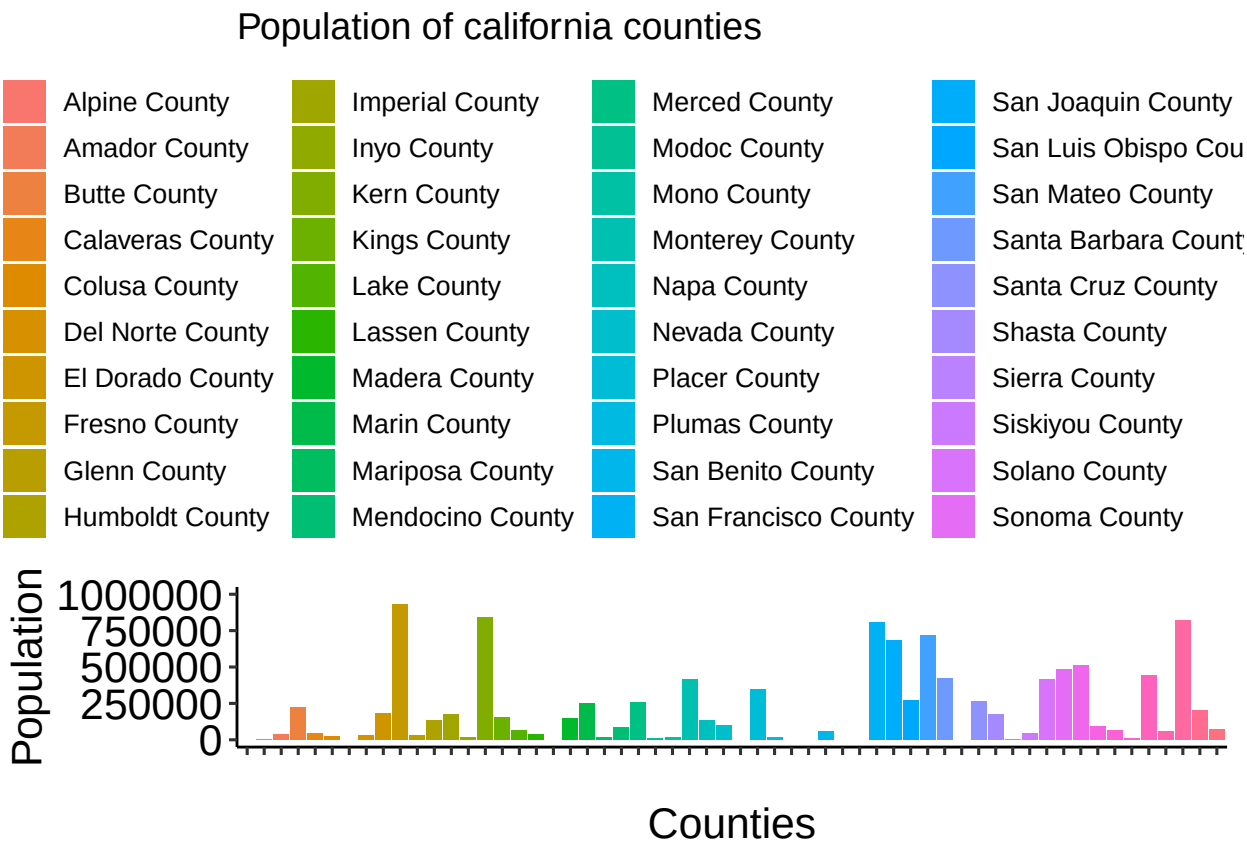
```
##            Id Id2  Geography  Target.Geo.Id Target.Geo.Id2
## 1 0400000US06   6 California 0500000US06001           6001
## 2 0400000US06   6 California 0500000US06003           6003
## 3 0400000US06   6 California 0500000US06005           6005
## 4 0400000US06   6 California 0500000US06007           6007
## 5 0400000US06   6 California 0500000US06009           6009
## 6 0400000US06   6 California 0500000US06011           6011
##                  Geographic.area           county Population Housing.units
## 1   California - Alameda County   Alameda County    1510271        582549
## 2    California - Alpine County    Alpine County       1175          1760
## 3    California - Amador County    Amador County      38091         18032
## 4     California - Butte County     Butte County     220000         95835
## 5 California - Calaveras County Calaveras County      45578         27925
## 6    California - Colusa County    Colusa County      21419          7883
##   Area.in.square.miles...Total.area Area.in.square.miles...Water.area
## 1                            821.33                             82.31
## 2                            743.18                              4.85
## 3                            605.96                             11.37
## 4                           1677.13                             40.67
## 5                           1036.93                             16.92
## 6                           1156.36                              5.63
##   Area.in.square.miles...Land.area
## 1                           739.02
## 2                           738.33
## 3                           594.58
## 4                          1636.46
## 5                          1020.01
## 6                          1150.73
##   Density.per.square.mile.of.land.area...Population
## 1                                             2043.6
## 2                                                1.6
## 3                                               64.1
## 4                                              134.4
## 5                                               44.7
## 6                                               18.6
##   Density.per.square.mile.of.land.area...Housing.units
## 1                                                788.3
## 2                                                  2.4
## 3                                                 30.3
## 4                                                 58.6
## 5                                                 27.4
## 6                                                  6.9
```

```
colnames(census)<- c("ID","ID2","Geog","Geoid1","Geoid2","Geographicarea",
                     "county","Population","Housing units","totalarea","waterarea",
                     "leandarea","popdens","housedens")
```

Population size per county

The following plot states the population size per county

```
## Warning: Removed 9 rows containing missing values (position_stack).
```



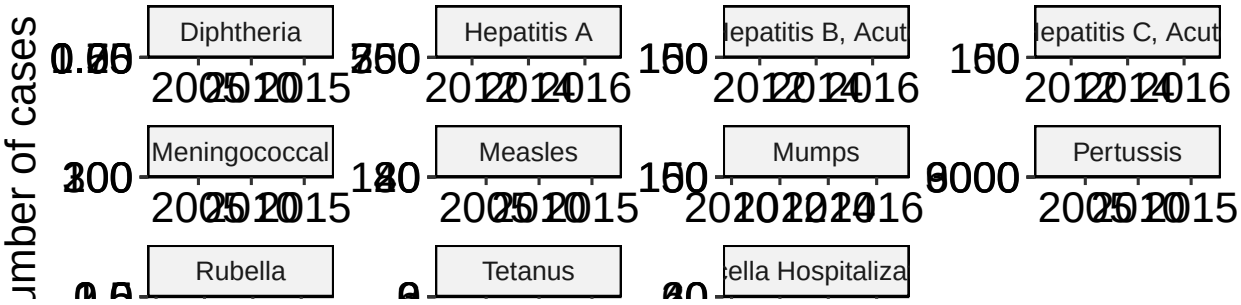
Disease distribution per county

The following plot states the population size per county

```
##      disease  county year count
## 1 Diphtheria Alameda 2001     0
## 2 Diphtheria Alameda 2002     0
## 3 Diphtheria Alameda 2003     0
## 4 Diphtheria Alameda 2004     0
## 5 Diphtheria Alameda 2005     0
## 6 Diphtheria Alameda 2006     0
```

```
##      disease     county 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
## 1 Diphtheria    Alameda    0    0    0    0    0    0    0    0    0    0
## 2 Diphtheria     Alpine    0    0    0    0    0    0    0    0    0    0
## 3 Diphtheria     Amador    0    0    0    0    0    0    0    0    0    0
## 4 Diphtheria      Butte    0    0    0    0    0    0    0    0    0    0
## 5 Diphtheria  Calaveras    0    0    0    0    0    0    0    0    0    0
## 6 Diphtheria California    0    1    0    0    0    0    0    0    0    0
##   2011 2012 2013 2014 2015 2016 2017
## 1    0    0    0    0    0    0    0
```

```
## 2    0    0    0    0    0    0    0
## 3    0    0    0    0    0    0    0
## 4    0    0    0    0    0    0    0
## 5    0    0    0    0    0    0    0
## 6    0    0    0    0    0    0    0
```



#####Data wrangling to select data for 2010 only as census data is for 2010 only

```
#split the data for county from census
head(cenus)
```

```
##            ID ID2       Geog         Geoid1 Geoid2
## 1 0400000US06   6 California 0500000US06001   6001
## 2 0400000US06   6 California 0500000US06003   6003
## 3 0400000US06   6 California 0500000US06005   6005
## 4 0400000US06   6 California 0500000US06007   6007
## 5 0400000US06   6 California 0500000US06009   6009
## 6 0400000US06   6 California 0500000US06011   6011
##                 Geographicarea           county Population Housing units
## 1   California - Alameda County   Alameda County    1510271        582549
## 2    California - Alpine County    Alpine County       1175          1760
## 3    California - Amador County    Amador County      38091         18032
## 4     California - Butte County     Butte County     220000         95835
## 5 California - Calaveras County Calaveras County      45578         27925
## 6    California - Colusa County    Colusa County      21419          7883
##   totalarea waterarea leandarea popdens housedens
```

```
## 1    821.33     82.31    739.02  2043.6     788.3
## 2    743.18      4.85    738.33     1.6       2.4
## 3    605.96     11.37    594.58    64.1      30.3
## 4   1677.13     40.67   1636.46   134.4      58.6
## 5   1036.93     16.92   1020.01    44.7      27.4
## 6   1156.36      5.63   1150.73    18.6       6.9
```

```
census<- cenus %>%
  na.omit() %>%
  separate("county",into=c("county","just "))
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 14 rows [7, 8,
## 9, 19, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44].
```

```
##merge by county
dem.dis<- full_join(census,dis,by="county")
```

```
## Warning: Column `county` joining character vector and factor, coercing into
## character vector
```

```
head(dem.dis)
```

```
##            ID ID2       Geog         Geoid1 Geoid2
## 1 0400000US06   6 California 0500000US06001   6001
## 2 0400000US06   6 California 0500000US06001   6001
## 3 0400000US06   6 California 0500000US06001   6001
## 4 0400000US06   6 California 0500000US06001   6001
## 5 0400000US06   6 California 0500000US06001   6001
## 6 0400000US06   6 California 0500000US06001   6001
##                Geographicarea  county  just  Population Housing units
## 1 California - Alameda County Alameda County    1510271        582549
## 2 California - Alameda County Alameda County    1510271        582549
## 3 California - Alameda County Alameda County    1510271        582549
## 4 California - Alameda County Alameda County    1510271        582549
## 5 California - Alameda County Alameda County    1510271        582549
## 6 California - Alameda County Alameda County    1510271        582549
##   totalarea waterarea leandarea popdens housedens    disease year count
## 1    821.33     82.31    739.02  2043.6     788.3 Diphtheria 2001     0
## 2    821.33     82.31    739.02  2043.6     788.3 Diphtheria 2002     0
## 3    821.33     82.31    739.02  2043.6     788.3 Diphtheria 2003     0
## 4    821.33     82.31    739.02  2043.6     788.3 Diphtheria 2004     0
## 5    821.33     82.31    739.02  2043.6     788.3 Diphtheria 2005     0
## 6    821.33     82.31    739.02  2043.6     788.3 Diphtheria 2006     0
```

```
tail(dem.dis)
```

```
##        ID ID2 Geog Geoid1 Geoid2 Geographicarea     county just
## 7679 <NA>  NA <NA>   <NA>     NA           <NA> Santa Cruz  <NA>
## 7680 <NA>  NA <NA>   <NA>     NA           <NA> Santa Cruz  <NA>
## 7681 <NA>  NA <NA>   <NA>     NA           <NA> Santa Cruz  <NA>
## 7682 <NA>  NA <NA>   <NA>     NA           <NA> Santa Cruz  <NA>
```

```
## 7683 <NA>  NA <NA>   <NA>     NA           <NA> Santa Cruz  <NA>
## 7684 <NA>  NA <NA>   <NA>     NA           <NA> Santa Cruz  <NA>
##      Population Housing units totalarea waterarea leandarea popdens
## 7679         NA             NA        NA        NA        NA      NA
## 7680         NA             NA        NA        NA        NA      NA
## 7681         NA             NA        NA        NA        NA      NA
## 7682         NA             NA        NA        NA        NA      NA
## 7683         NA             NA        NA        NA        NA      NA
## 7684         NA             NA        NA        NA        NA      NA
##      housedens                    disease year count
## 7679        NA Varicella Hospitalizations 2012     0
## 7680        NA Varicella Hospitalizations 2013     0
## 7681        NA Varicella Hospitalizations 2014     0
## 7682        NA Varicella Hospitalizations 2015     1
## 7683        NA Varicella Hospitalizations 2016     0
## 7684        NA Varicella Hospitalizations 2017     0
```

##data only for pertussis as it is the most common disease

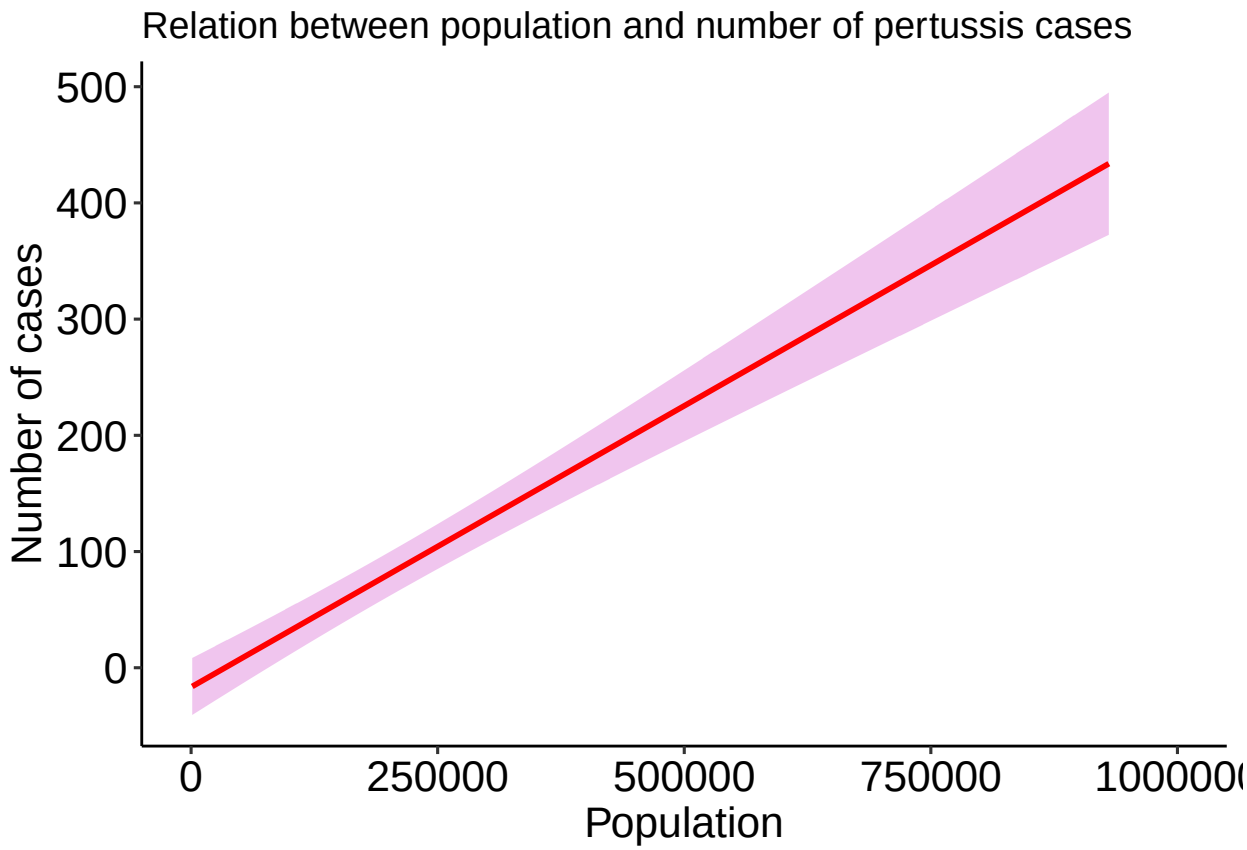
```
###does the cases depend on the population
dem.dis2<-dem.dis %>% filter(year==2010) %>% group_by(disease) %>%
  select(county,disease,Population,count,popdens)
###pertusis
de.dis.2010.per<-dem.dis %>%
  filter(year==2010) %>%
  select(county,year,Population,disease,count,popdens) %>%
 filter(disease== "Pertussis")
```

###GGplot with regression line to understand trenad with #A. Population

```
### relation between population and cases of pertusis
ggplot(de.dis.2010.per, aes(Population,count)) +
  geom_smooth(method=lm, forumla =count~Population,se=T,col="red",fill="orchid")+
  xlim(0,1000000)+
  ggpubr::theme_pubr()+
  xlab("Population")+
  ylab("Number of cases")+
  theme(axis.text.x = element_text(size=16))+
  theme(axis.title.x = element_text(size=16))+
  theme(axis.text.y = element_text(size=16))+
  theme(axis.title.y = element_text(size=16))+
  labs(title="Relation between population and number of pertussis cases ")
```

```
## Warning: Ignoring unknown parameters: forumla
```

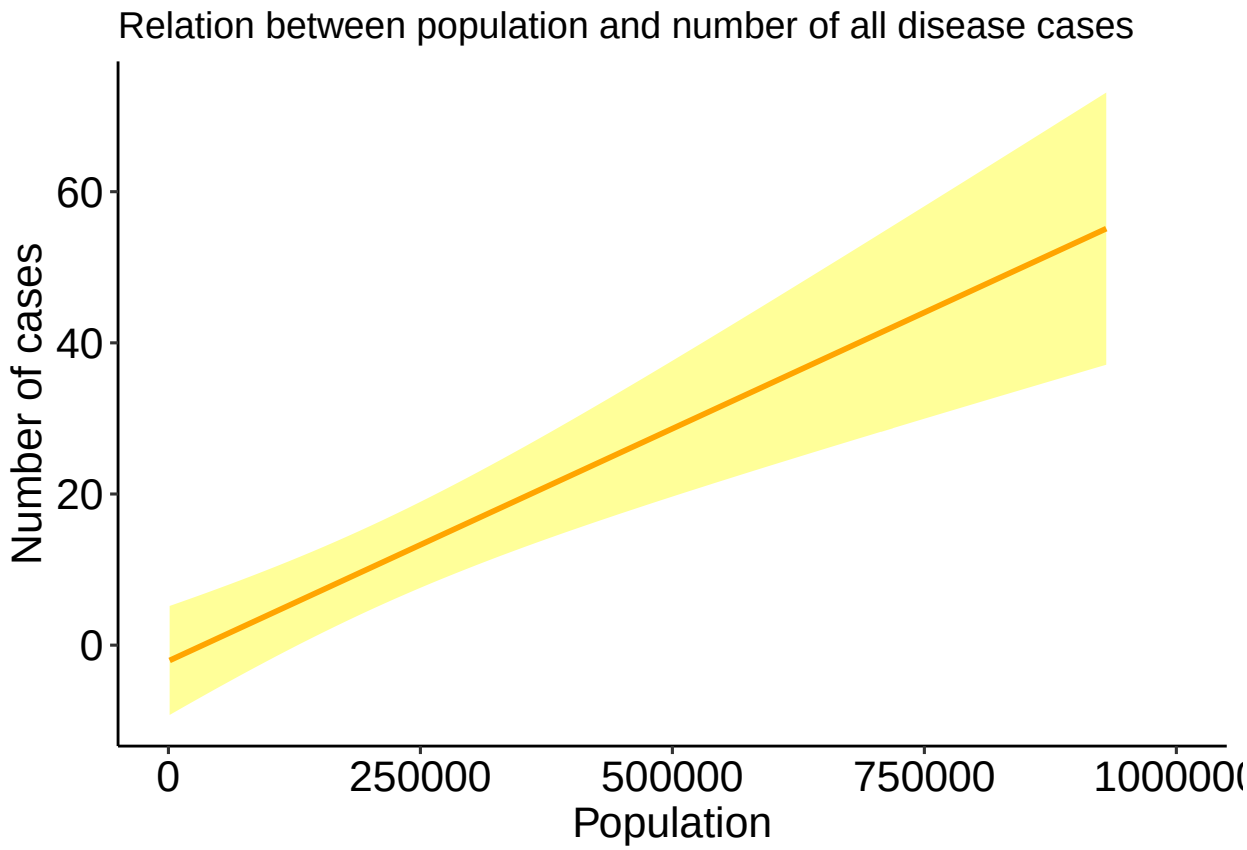
```
## Warning: Removed 19 rows containing non-finite values (stat_smooth).
```




```
### relation between population and cases of all diseases
ggplot(dem.dis2, aes(Population,count)) +
  geom_smooth(method=lm, forumla =count~Population,se=T,col="orange",fill="yellow")+
  xlim(0,1000000)+
  ggpubr::theme_pubr()+
  xlab("Population")+
  ylab("Number of cases")+
  theme(axis.text.x = element_text(size=16))+
  theme(axis.title.x = element_text(size=16))+
  theme(axis.text.y = element_text(size=16))+
  theme(axis.title.y = element_text(size=16))+
  labs(title="Relation between population and number of all disease cases ")
```

```
## Warning: Ignoring unknown parameters: forumla
```

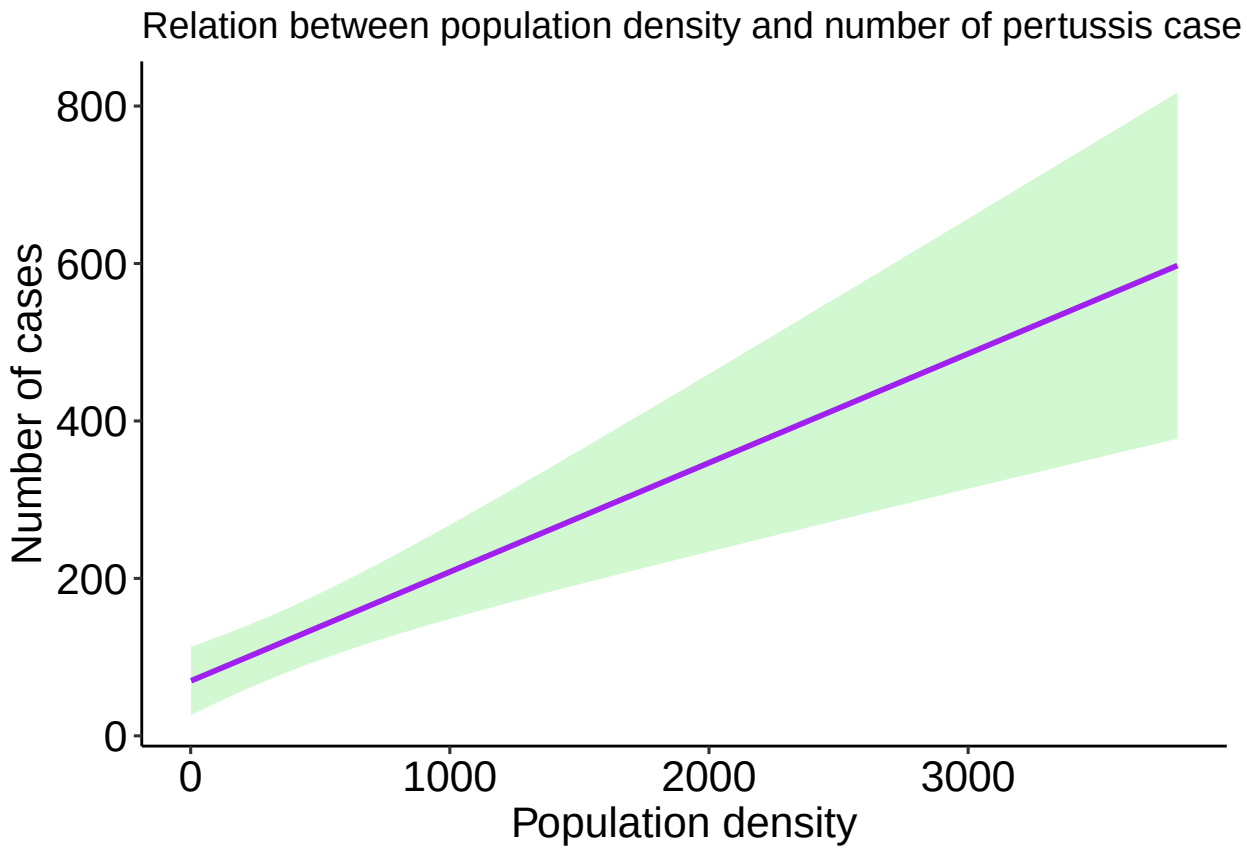
```
## Warning: Removed 152 rows containing non-finite values (stat_smooth).
```




```
#B. Density
####################densiyt
### relation between population and cases of pertusis
ggplot(de.dis.2010.per, aes(popdens,count)) +
  geom_smooth(method=lm, forumla =count~popdens,se=T,col="purple",fill="light green")+
  #xlim(0,1000000)+
  ggpubr::theme_pubr()+
  xlab("Population density")+
  ylab("Number of cases")+
  theme(axis.text.x = element_text(size=16))+
  theme(axis.title.x = element_text(size=16))+
  theme(axis.text.y = element_text(size=16))+
  theme(axis.title.y = element_text(size=16))+
  labs(title="Relation between population density and number of pertussis cases ")
```

```
## Warning: Ignoring unknown parameters: forumla
```

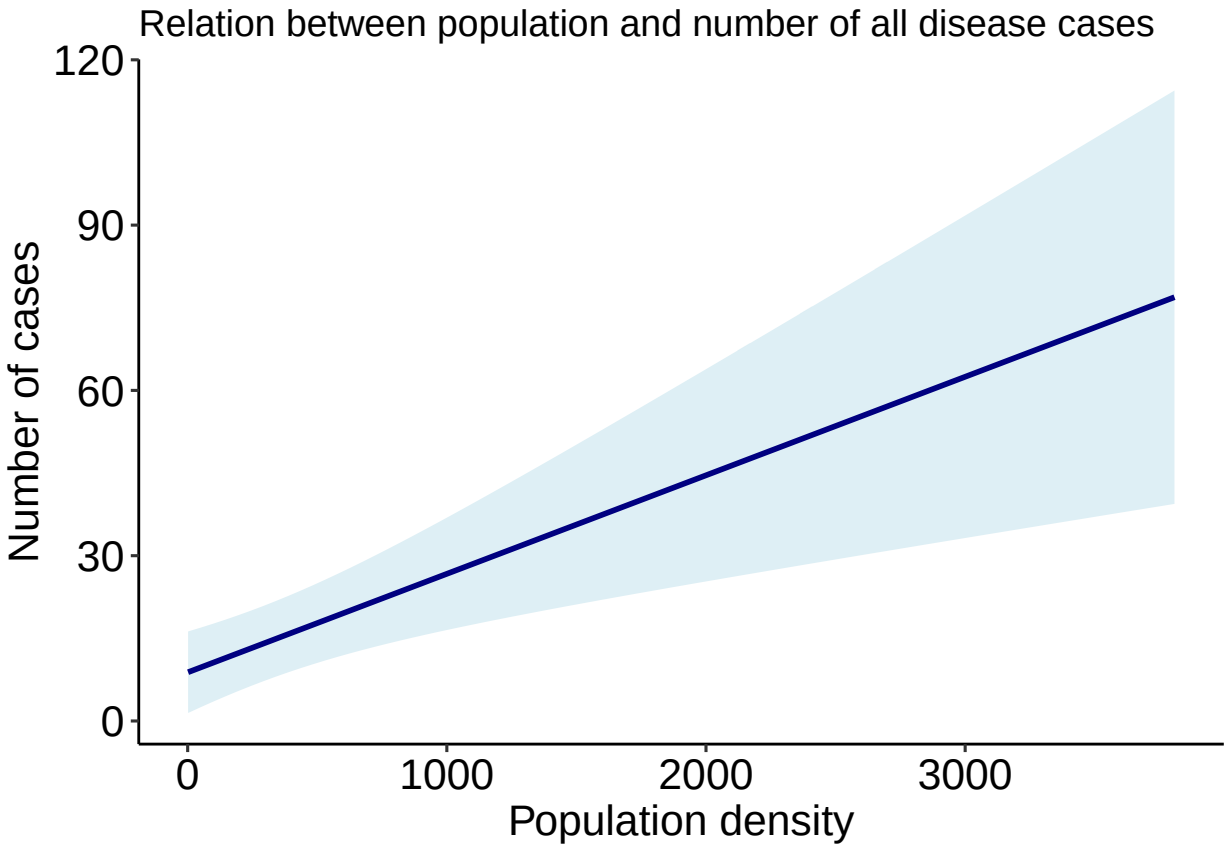
```
## Warning: Removed 15 rows containing non-finite values (stat_smooth).
```




```
### relation between population and cases of all diseases
ggplot(dem.dis2, aes(popdens,count)) +
  geom_smooth(method=lm, forumla =count~podens,se=T,col="navy blue",fill="ligh tblue")+
  #xlim(0,1000000)+
  ggpubr::theme_pubr()+
  xlab("Population density")+
  ylab("Number of cases")+
  theme(axis.text.x = element_text(size=16))+
  theme(axis.title.x = element_text(size=16))+
  theme(axis.text.y = element_text(size=16))+
  theme(axis.title.y = element_text(size=16))+
  labs(title="Relation between population and number of all disease cases ")
```

```
## Warning: Ignoring unknown parameters: forumla
```

```
## Warning: Removed 120 rows containing non-finite values (stat_smooth).
```


Relation between population and number of all disease cases
120
90
60
30
0
Number of cases
0
1000
2000
3000
Population density