MMPH6117 Assignment 1

Association of invasive meningococcal disease and environmental and virologic exposure

Name: Mingze Gao   UID: 3035426241

Weekly counts of invasive meningococcal disease (IMD) reported cases from country A in 1997-2000 were collected and analyzed, to study associations between IMD and environmental and virologic exposure. The file 'IMD.csv' stored the following variables which were found to be significant factors from the literature in other places:
(1) flu: influenza isolation rate from laboratory
(2) maxtemp: weekly maximum temperature
(3) IMD: weekly number of IMD cases
(4) pop: population size

```
library(readr)
IMD <- read_csv("../Data/CMED6020 Assignment1 - IMD.CSV")
```

```
##
## -- Column specification --------------------------------------------------
## cols(
##   flu = col_double(),
##   maxtemp = col_double(),
##   imd = col_double(),
##   pop = col_double()
## )
```

a) Fit a poisson regression model to predict weekly IMD cases using flu and maxtemp, accounting for the increasing population size in 1997-2000. Summarize your results in a table. [4 marks]

```
summary(model_poi <- glm(imd ~ flu + maxtemp + offset(log10(pop)), data = IMD, family = poisson()))
```

```
##
## Call:
## glm(formula = imd ~ flu + maxtemp + offset(log10(pop)), family = poisson(),
##     data = IMD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.420   -7.301   -4.463    3.566   36.169
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.463795   0.021992 -157.51  < 2e-16 ***
## flu          1.448713   0.101383   14.29  < 2e-16 ***
## maxtemp      0.011365   0.002078    5.47  4.5e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 20710  on 207  degrees of freedom
## Residual deviance: 20447  on 205  degrees of freedom
## AIC: 21774
##
## Number of Fisher Scoring iterations: 5
```

b) Assess the goodness of fit of the model. [1 marks]

```
deviance(model_poi) / df.residual(model_poi)
```

```
## [1] 99.74293
```

The residual deviance / df is way larger than 1, therefore this model is poorly fitted.

c) Quote the mean and variance of the weekly IMD cases. Is there any evidence of overdispersion? [1 marks]

```
imd_mean <- mean(IMD$imd)
imd_var <- var(IMD$imd)
print (paste("Mean:", imd_mean, sep = " "))
```

```
## [1] "Mean: 137.043269230769"
```

```
print (paste("Var:", imd_var, sep = " "))
```

```
## [1] "Var: 16868.5536742847"
```

The difference between mean and variance of the weekly IMD cases are extremely high. Possion distribution ideally have equal parameters. So, yes evidence of overdispersion.

d) Fit a negative binomial regression model to predict weekly IMD cases and summarize your results in a table. [4 marks]

```
require(MASS)
```

```
## Loading required package: MASS
```

```
summary(model_nb <- glm.nb(imd ~ flu + maxtemp + offset(log10(pop)), data = IMD))
```

```
##
## Call:
## glm.nb(formula = imd ~ flu + maxtemp + offset(log10(pop)), data = IMD,
##     init.theta = 1.392209478, link = log)
##
## Deviance Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.9030  -0.8229  -0.4881   0.3147   2.7824
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.48902    0.21319 -16.366   <2e-16 ***
## flu          1.54678    1.00604   1.537    0.124
## maxtemp      0.01294    0.02074   0.624    0.533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.3922) family taken to be 1)
##
##     Null deviance: 234.03  on 207  degrees of freedom
## Residual deviance: 231.18  on 205  degrees of freedom
## AIC: 2458
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  1.392
##          Std. Err.:  0.126
##
##  2 x log-likelihood:  -2449.964
```

e) Assess the goodness of fit of the model. [1 marks]

```
deviance(model_nb) / df.residual(model_nb)
```

```
## [1] 1.127713
```

The model fit the data satisfactorily.

f) Calculate the AIC for the poisson and negative binomial regression models and select the best model based on AIC and goodness of fit. [1 marks]

```
print (paste("aic_poi:", model_poi$aic, sep = " "))
```

```
## [1] "aic_poi: 21773.8442659117"
```

```
print (paste("aic_nb:", model_nb$aic, sep = " "))
```

```
## [1] "aic_nb: 2457.96381241714"
```

Based on AIC and goodness of fit, negative binomial regreesion model should be chosen, as AIC is lower and deviance not far from 1.

g) Assess if the linear effects for the variables flu and maxtemp are adequate. [2 marks]

```
model_nb$coefficients
```

```
## (Intercept)        flu     maxtemp
##   -3.489018    1.546777    0.012937
```

The coefficient for flu and maxtemp are 1.534 and 0.013 respectively. Linear effects for flu is adequate whilst not much adequate for maxtemp.

h) Assess if there is any collinearity problem in your final model? [2 marks]

```
require(car)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```
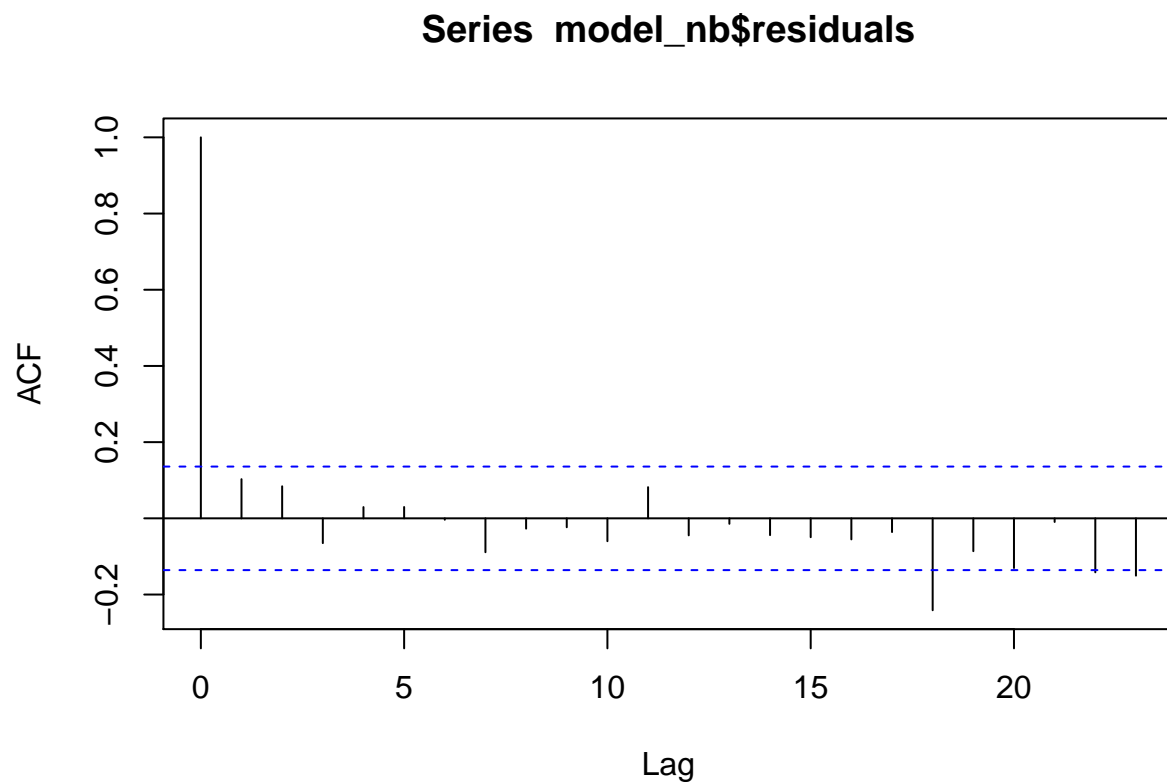
```
vif(model_nb)
```
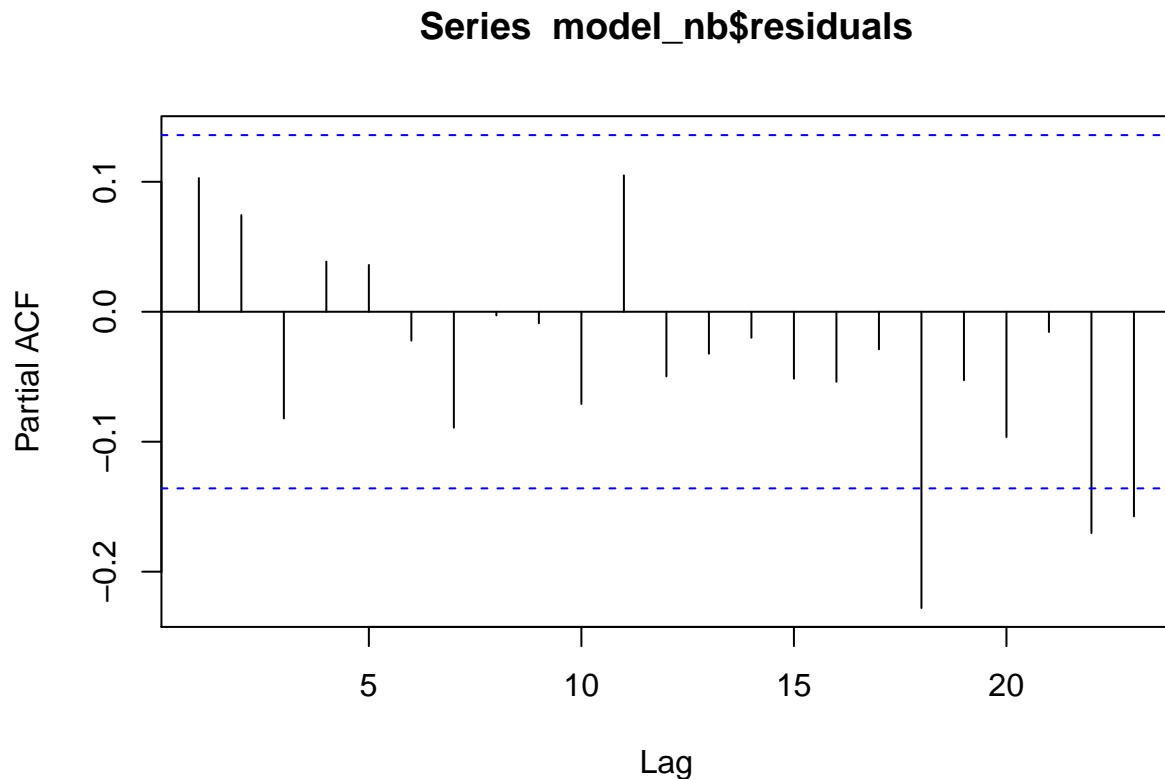
```
##      flu  maxtemp
## 1.019899 1.019899
```

The VIF is close to 1 which indicates no clear multi-collinearity problem between variables in the model.

i) Assess if there is any unexplained serial correlation after fitting your final model. [2 marks]

```
acf(model_nb$residuals)
```

Series model_nb$residuals

```
pacf(model_nb$residuals)
```

Series model_nb$residuals

ACF and partial ACF plots has not shown exceeding confidence limit in the first few lags so there should be no unexplained serial correlation.

j) Draw an overall conclusion on the findings. [3 marks]

```
round(exp(cbind(coef(model_nb), confint(model_nb))),3)
```

```
## Waiting for profiling to be done...

##                 2.5 % 97.5 %
## (Intercept) 0.031 0.020  0.047
## flu         4.696 0.635 34.710
## maxtemp     1.013 0.973  1.054
```

We can present the weekly counts of IMD for flu, which is 4.696 (95% CI: 0.635 - 34.710). We can say that the weekly counts of IMD varies a great deal between different time. The fit of negative binominal model seems to be acceptable compare with possion regression model.

k) Based on the final model, predict the number of IMD cases in a week, if the influenza isolation rate is 0.3, with a maximum temperature of 5°C and a population of 110 million. [2 marks]

```r
new_data <- data.frame(flu = 0.3, maxtemp = 5, pop = 110000000)
predict(model_nb, newdata = new_data, type = "response")
```

```
##        1
## 160.9499
```

l) Based on the final model, which of the conditions below will have a higher population risk of IMD, so
   more attention should be paid to suspected IMD cases for earlier treatment? [2 marks]

1. Influenza isolation rate = 0.1, maximum temperature = 10ºC
2. Influenza isolation rate = 0.4, maximum temperature = -5ºC

```r
new_data <- data.frame(flu = c(0.1, 0.4), maxtemp = c(10, -5), pop = 110000000)
predict(model_nb, newdata = new_data, type = "response")
```

```
##        1        2
## 126.0177 165.0751
```

Condition 2 have a higher population resk of IMD.