Assignment 1 suggested solution

(The R script is for reference only)

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]

Table 1: Relative risks (RR) and the corresponding 95% confidence interval (CI) of the IMD

Variables	RR (95% CI)	p-value
Temperature	1.011 (1.007, 1.016)	<0.001
Influenza isolation rate	4.25 (3.49, 5.19)	<0.001

Influenza isolation rate is significantly associated (p<0.001) with the increase in IMD (RR=5.01). Temperature is also significantly associated (p<0.001) with IMD at 5% level of significance.

```
data <- read.csv(file="imd.csv")
summary(data)
pois.imd <- glm(imd~offset(log(pop))+flu+maxtemp,
data=data,family=poisson)
summary(pois.imd)
round(exp(cbind(coef(pois.imd),confint(pois.imd))),3)
```

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

Since the deviance/df (Value/df=99.74) is much greater than 1, it indicated severe lack of fit for the model.

```
deviance(pois.imd)/df.residual(pois.imd)
```

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

The mean and variance of the weekly IMD cases are 137 and 16869, respectively. The variance of IMD cases is much larger than the mean which indicates overdispersion.

```
mean(data$imd)
var(data$imd)
```

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

Table 2: Relative risks (RR) and the corresponding 95% confidence interval (CI) of the IMD

Variables	**RR (95% CI)**	**p-value**
Temperature	1.01 (0.97, 1.05)	0.533
Influenza isolation rate	4.69 (0.63, 34.65)	0.125

We re-fitted the model with family of negative binomial. Influenza isolation seemed to be positively associated with IMD. However, both the effects of influenza isolation rate ($p=0.125 > 0.05$) and temperature ($p=0.533 > 0.05$) were not statistically significant.

```
require(MASS)
nb.imd <- glm.nb(imd~offset(log(pop))+flu+maxtemp,data=data)
summary(nb.imd)
round(exp(cbind(coef(nb.imd),confint(nb.imd))),2)
```

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

The deviance to df ratio reduced to 1.13 which is just slightly greater than 1 (also overdispersion parameter = 1/1.392 = 0. 718). The negative binomial regression has a satisfactory fit and is more appropriate than poisson regression.

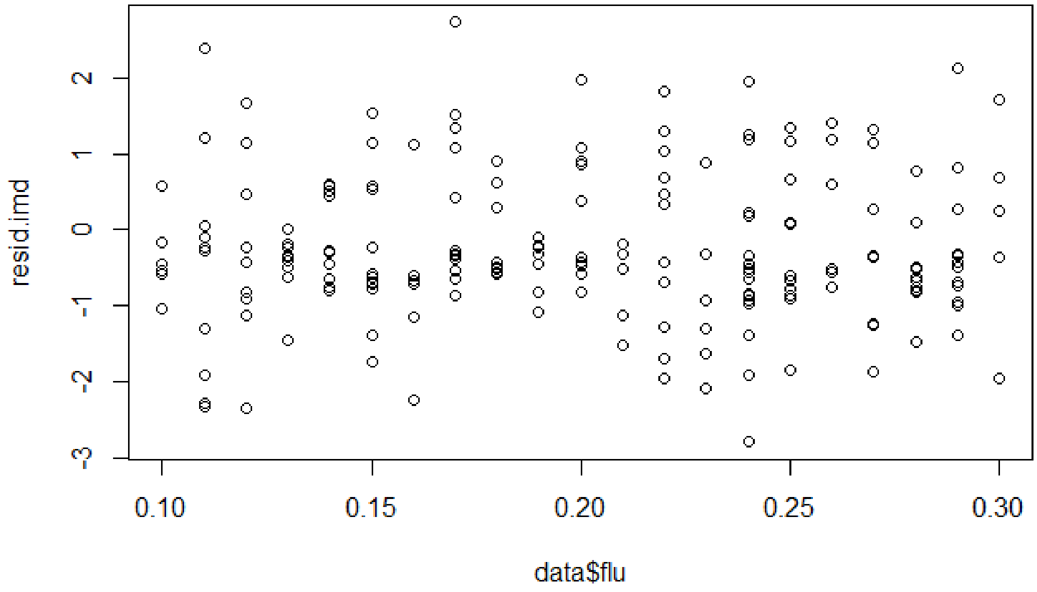
```
deviance(nb.imd)/df.residual(nb.imd)
```

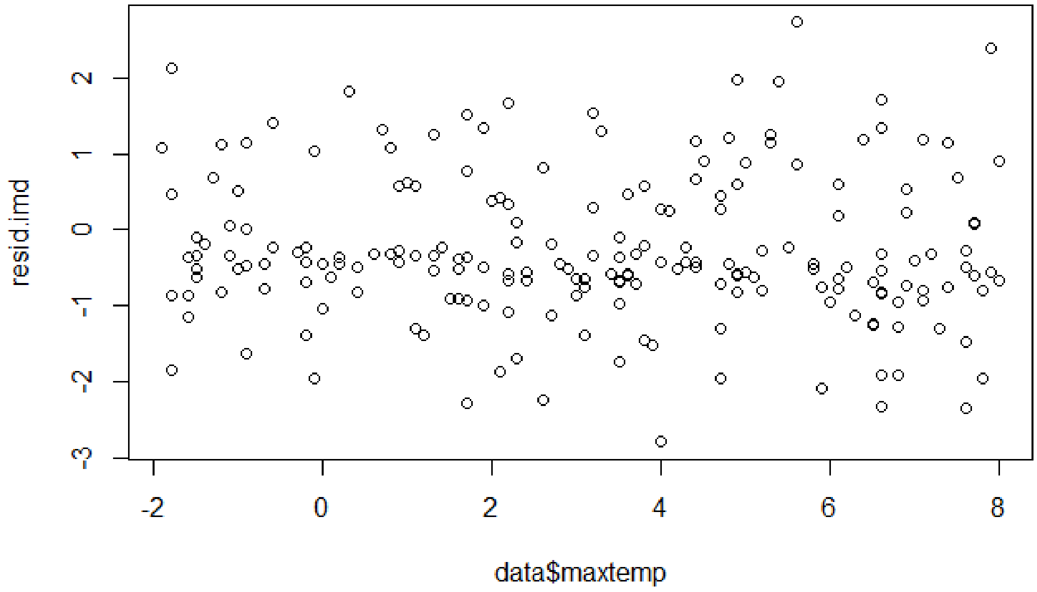
f) Calculate the AIC for the poisson and negative binomial regression models and select the best model based on AIC and goodness of fit. [1 marks]

The AIC for the Poisson and negative binomial regression models are 21773 and 2458 respectively. A lower AIC indicates a ‘better’ model. So based on the results of AIC and goodness of fit, the negative binomial regression model is preferred.

```
AIC(pois.imd,nb.imd)
```

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





Consider the residual plots, the linear effects for the variables *flu* and *maxtemp* are adequate.

```
resid.imd<-rstudent(nb.imd)

par(mfrow=c(2,1))
plot(data$flu, resid.imd)
plot(data$maxtemp, resid.imd)
```

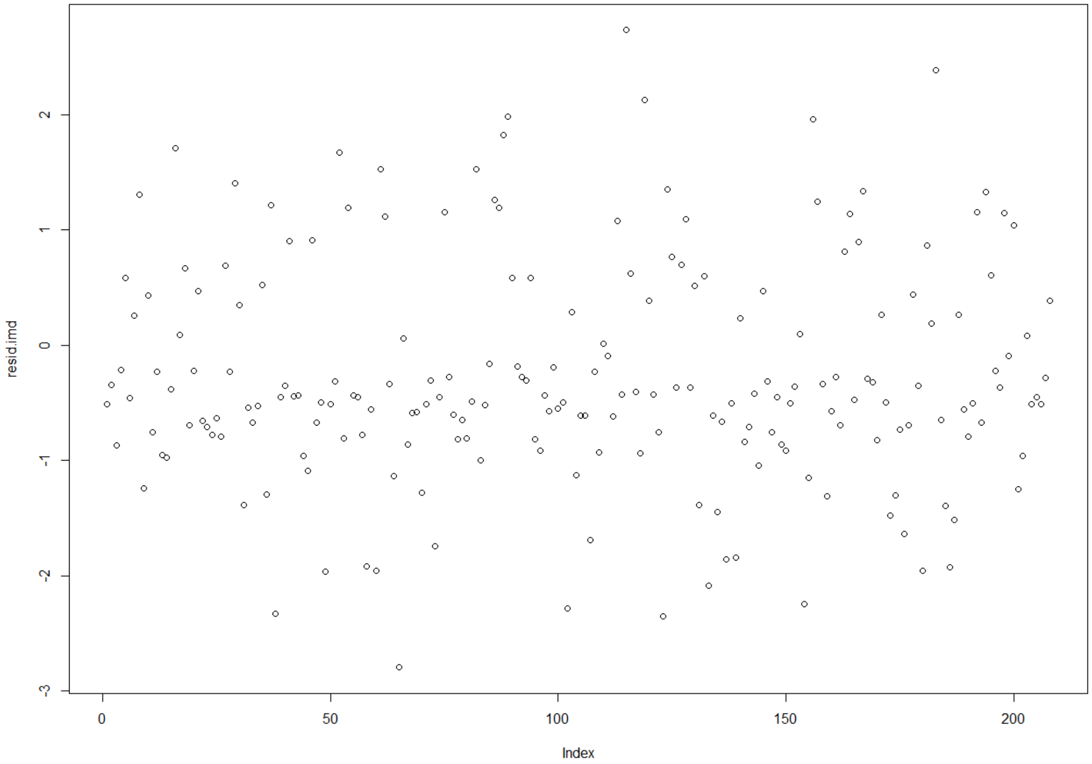
h) Assess if there is any collinearity problem in your final model? [2 marks]

The VIF for flu isolation rate and temperature are 1.02. There is no collinearity problem in the model.

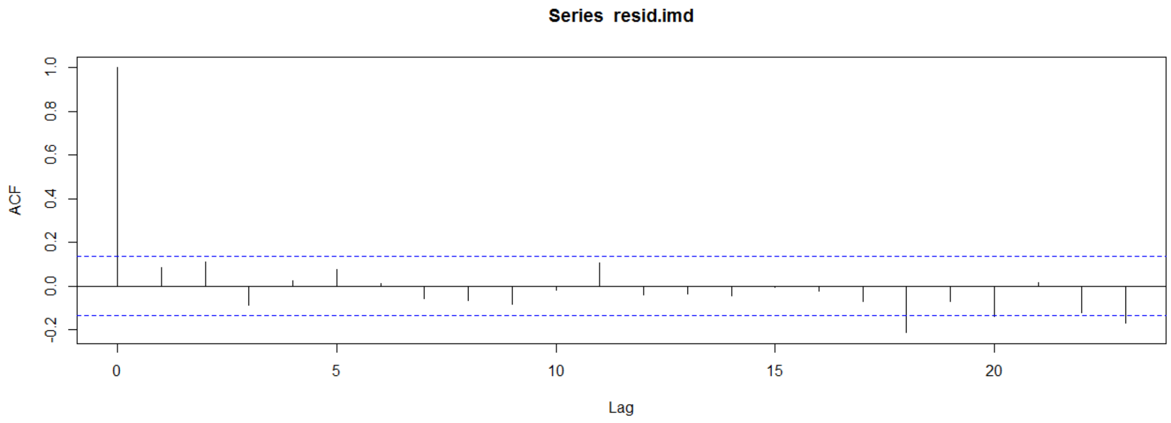
```
require(car)
vif(nb.imd)
```

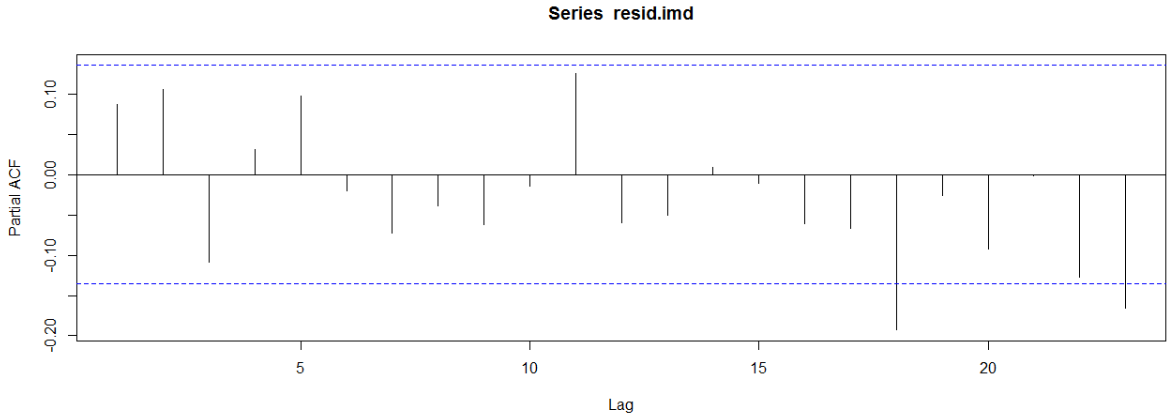
i) Assess if there is any unexplained serial correlation after fitting your final model. [2 marks]

Deviance residuals against week or ACF / partial ACF plots do not indicate any serial correlation.



[optional: ACF / PACF]





```
plot(resid.imd)
resid.imd<-rstudent(nb.imd)
par(mfrow=c(2,1))
acf(resid.imd)
pacf(resid.imd)
```

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

The final model is $log(E(Y)) = \log(pop) - 13.97 + 1.54 \cdot flu + 0.01 \cdot maxtemp$

So based on the results of AIC and goodness of fit, the negative binomial regression model is selected. We found no significant association between IMD cases with temperature or influenza activity.

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 5ºC and a population of 110 million. [3 marks]

The predicted number of IMD cases is 159.

```
newdata <- data.frame(flu=0.3,maxtemp=5, pop=110*1e6)
```

```
predict(nb.imd, newdata, type="response")
```

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

The predicted incidence rates are 113 and 148 per 100 million for conditions 1 and 2 respectively Therefore, condition 2 has a higher risk of IMD and more attention should be paid to suspected IMD cases during periods with similar conditions.

```
pop.size <- 1e8 # just assume a population size

newdata.cond <- data.frame(flu=c (0.1,0.4),maxtemp=c(10,-5),
pop=pop.size)
inc.cond <- round(predict(nb.imd, newdata.cond,
type="response")/pop.size*1e8,0)

or alternatively
```

Based on the final model, the predicted risk is given by $E(Y)/pop = \exp(-13.97 + 1.54 \cdot flu + 0.01 \cdot maxtemp$

```
exp(sum(coef(nb.imd)*c(1,0.1,10)))*1e8
exp(sum(coef(nb.imd)*c(1,0.4,-5)))*1e8
```