STATS 330 Assignment 1

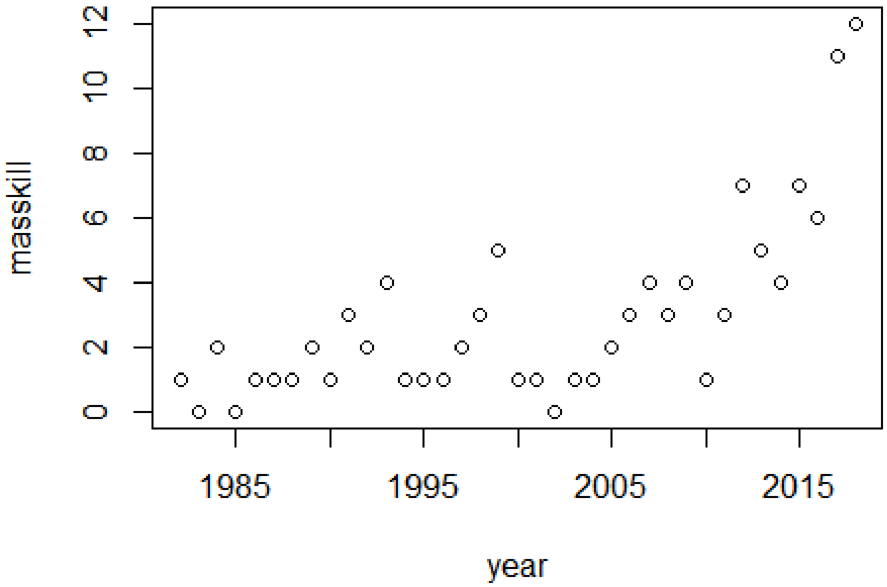
Sooyong Choi 915726645

Due Date: 12 noon Friday 16th August

```
setwd("C:/Users/rick9/OneDrive/Documents/Stats 330 A1")
masskill_file <- read.csv("masskill.csv")
```

PLot the number of mass killing incdidents per year over this period of time. Comment briefly

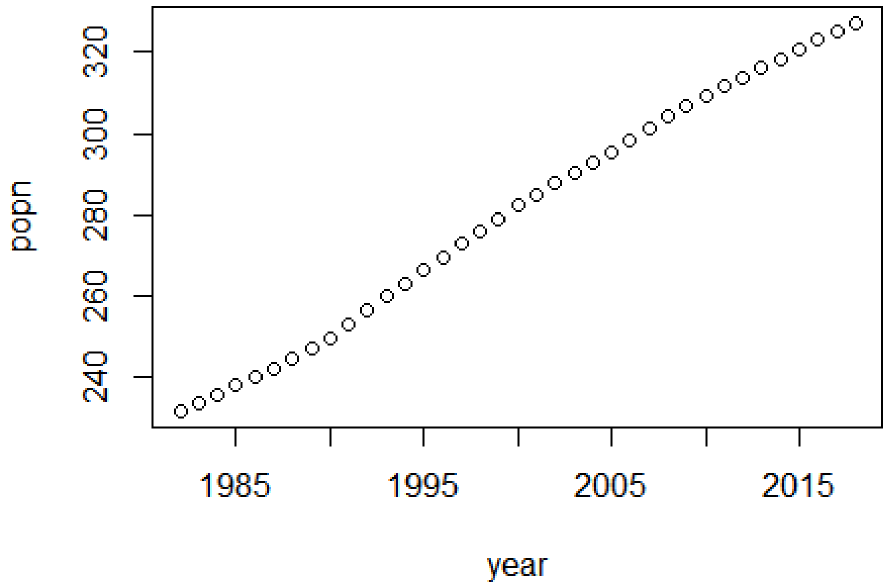
```
plot(masskill ~ year, data = masskill_file)
```



We can see that the number of mass killing incidenst per year over the period of year is increasing. The highest mass killing incident per year is 12 the highest around 2018

Plot the population of the USA over this period of time. Comment briefly on how the population is changing over this period of time.

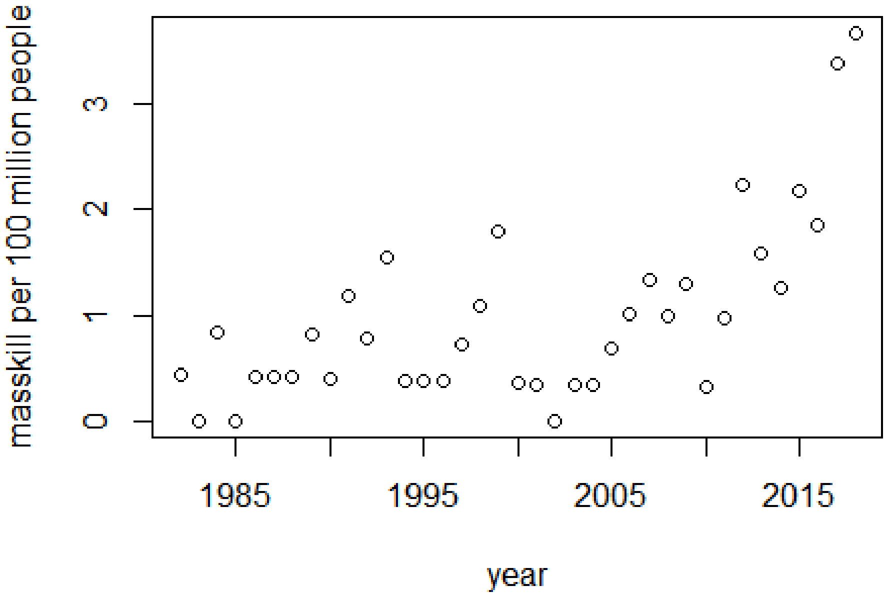
```
plot(popn ~ year, data = masskill_file)
```



We can see that the population of the USA (millions) is linearly increasing as time passes by. THe population ranges from around 230 (millions) to 330 (millions) and the year ranges from around 1980 to around 2018

Make a plot that shows the number of mass killing per year, per 100 million people. Comment briefly.

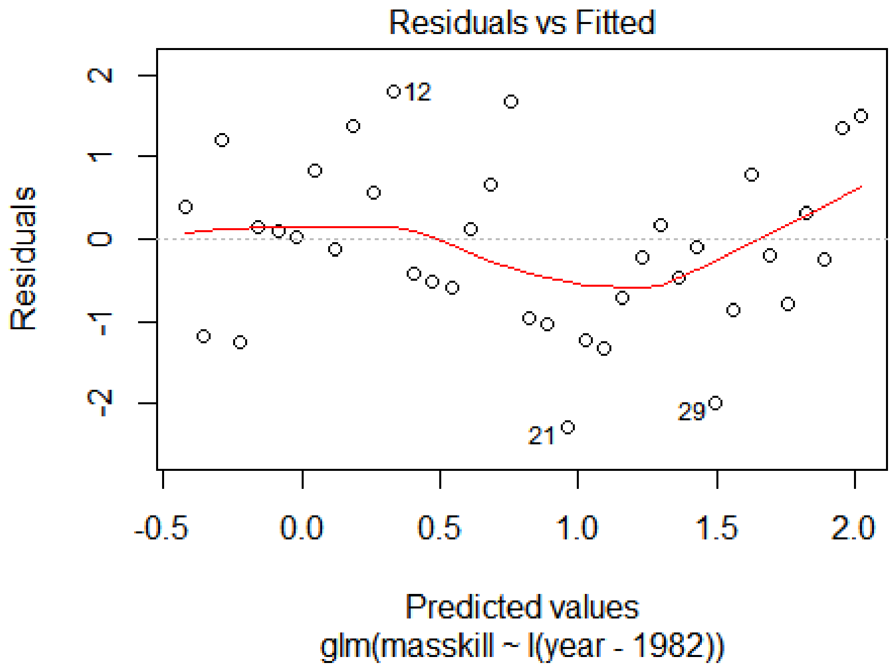
```
plot(masskill/(popn/100) ~ year, data = masskill_file,
     ylab = "masskill per 100 million people")
```



We can see that the masskill per 100 million people is overall increasing. There is also increase in scatter from 1980 to 2018. We can also note that at around year 2015 to 2018, there is a high scatter in masskill per 100 million people

Fit a 'linear model' in this model for the mass killing count, starting from 1982 as year 0, with the offset population exposure variable as follows. Comment, briefly on what you conclude from this output.

```
lin_poisson.fit <- glm(masskill ~ I(year - 1982), family = "poisson", offset
= log(popn/100), data = masskill_file)
plot(lin_poisson.fit, which = 1)
```



```
summary(lin_poisson.fit)

## 
## Call:
## glm(formula = masskill ~ I(year - 1982), family = "poisson", 
##     data = masskill_file, offset = log(popn/100))
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2838  -0.7970  -0.1192   0.5704   1.7945  
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -1.26445    0.28020  -4.513 6.40e-06 ***
## I(year - 1982)  0.05832    0.01051   5.550 2.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 71.466  on 36  degrees of freedom
## Residual deviance: 36.443  on 35  degrees of freedom
## AIC: 134.35
## 
## Number of Fisher Scoring iterations: 5

1 - pchisq(36.443, 35)
```

```
## [1] 0.4014187

exp(confint(lin_poisson.fit))

## Waiting for profiling to be done...

##                    2.5 %    97.5 %
## (Intercept)    0.1582762 0.4757867
## I(year - 1982) 1.0389748 1.0827558
```

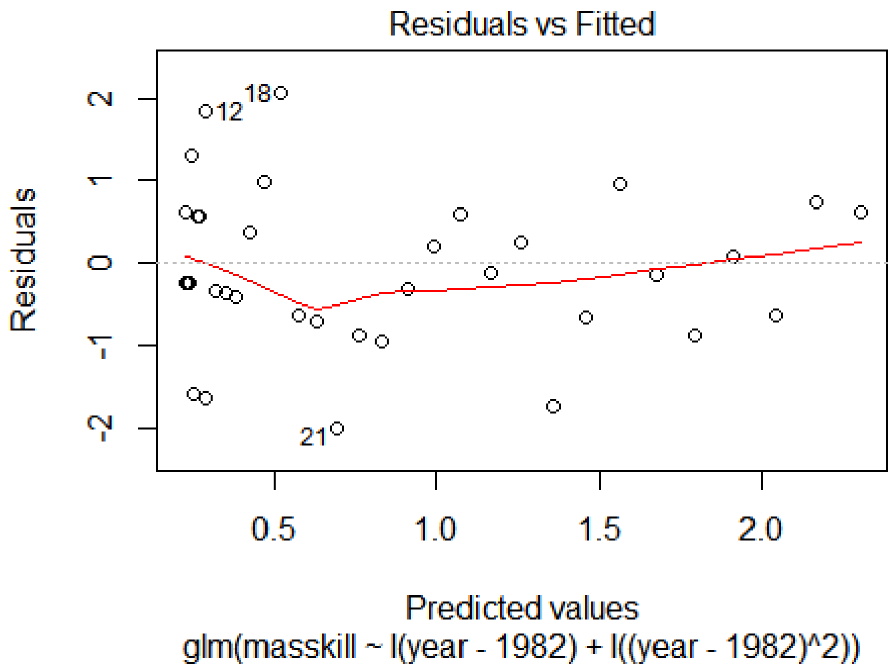
The residual plot shows a fairly constant scatter but presents a negative bias aroud the value between 1 to 1.5. There is a clear linear relationship between year and masskill (p - value close to 0) The test shows that the p value is large so we have no concern about the Poisson model. We estimate that the mean number of mass killings per year per 100 million people to increase between 104% and 108% for every year.

State, mathematically, the model for the count of mass killings that you are fitting here

e) $log(\mu_i) = \hat{\beta}_0 + \hat{\beta}_1 \times year_i + log(popn_i/100)$

Include an additional quadratic term in a new model model in this model for the mass killing count, starting from 1982 as year 0 with the offset popiulation exposure variable as follows: Comment, briefly, on what you conclude form this output.

```
quad_poisson.fit <- glm(masskill ~ I(year - 1982) + I((year - 1982)^2),
family = "poisson", offset = log(popn/100), data = masskill_file)
plot(quad_poisson.fit, which = 1)
```



```
summary(quad_poisson.fit)

##
## Call:
## glm(formula = masskill ~ I(year - 1982) + I((year - 1982)^2),
##     family = "poisson", data = masskill_file, offset = log(popn/100))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0022  -0.6369  -0.2376   0.5628   2.0609
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.5206165  0.3827536  -1.360   0.1738
## I(year - 1982)     -0.0388638  0.0414856  -0.937   0.3489
## I((year - 1982)^2)  0.0023419  0.0009919   2.361   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 71.466  on 36  degrees of freedom
## Residual deviance: 31.142  on 34  degrees of freedom
## AIC: 131.05
##
## Number of Fisher Scoring iterations: 5
```

```
1 - pchisq(31.142, 34)

## [1] 0.6084442

exp(confint(quad_poisson.fit))

## Waiting for profiling to be done...

##                         2.5 %   97.5 %
## (Intercept)         0.2622946 1.186055
## I(year - 1982)      0.8889913 1.046729
## I((year - 1982)^2) 1.0003567 1.004266
```

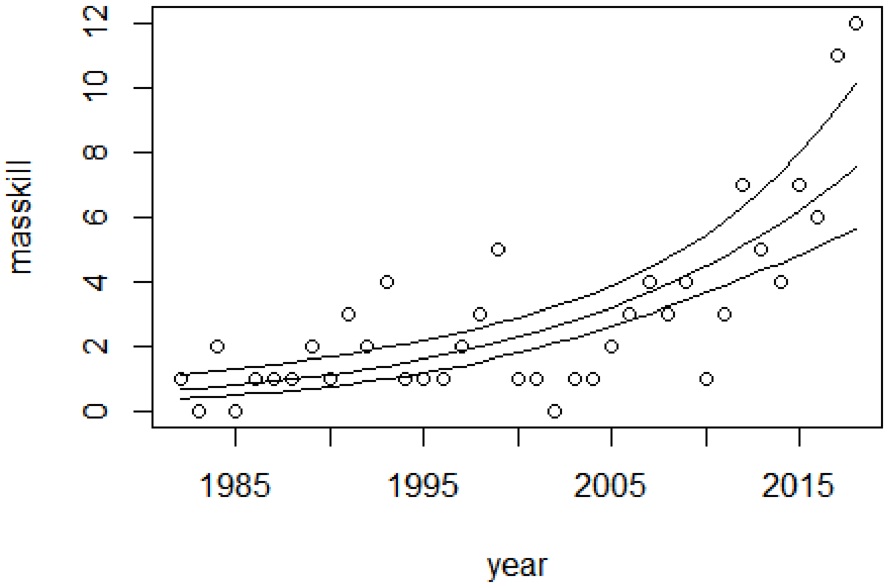
The residual plot shows slightly decrease in scatter as the value increases but overall it is patternless so it is not big of a concern. We also see that the quadratic term is statistically significant (p - value = 0.0182). The goodness-of-fit test shows that the p value is large so we have no concern about the Poisson model.

State, mathetmatically, the model for the count of mass killings that you are fitting here

g) $log(\mu_i) = \hat{\beta}_0 + \hat{\beta}_1 \times year_i + \hat{\beta}_2 \times year_i^2 + log(popn_i/100)$

Plot the data again with the linear model's expected counts superimposed along with the 95% confidence interval band for these band for these expected values. Comment, briefly, on how well your model fits these data.

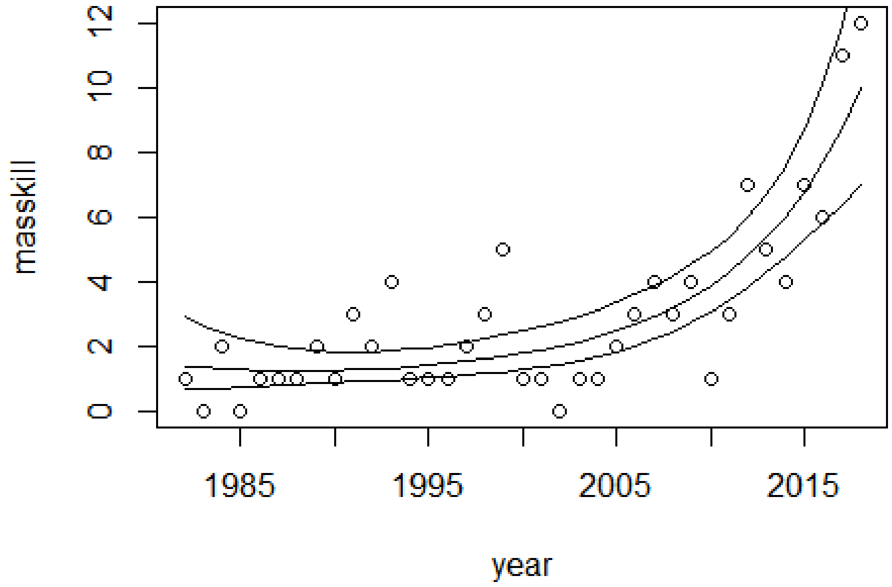
```
plot(masskill ~ year, data = masskill_file)
lin_pred <- predict(lin_poisson.fit, masskill_file, se.fit = TRUE)
lines(x=1982:2018, exp(lin_pred$fit))
lin_lower_bound <- lin_pred$fit - lin_pred$se.fit * 1.96
lin_higher_bound <-lin_pred$fit + lin_pred$se.fit * 1.96
lines(x=1982:2018, exp(lin_lower_bound))
lines(x=1982:2018, exp(lin_higher_bound))
```



The mean number of mass killing per year per 100 million people seems to be increasing steadily throughout the year. The linear poisson model with 95% confidence interval covers the majority of the data so the model represents the data well.

Plot the data again with the quadratic model's expected coutns superimposed along with the 95% confidence interval band for these expected values. Comment, briefly, on how well your model fits these data.

```r
plot(masskill ~ year, data = masskill_file)
quad_predict <- predict(quad_poisson.fit, masskill_file, se.fit = TRUE)
lines(x=1982:2018, exp(quad_predict$fit))
quad_lower_bound <- quad_predict$fit - quad_predict$se.fit * 1.96
quad_higher_bound <- quad_predict$fit + quad_predict$se.fit * 1.96
lines(x=1982:2018, exp(quad_lower_bound))
lines(x=1982:2018, exp(quad_higher_bound))
```



The mean number of mass kill per year per 100 million people seems to be increasing steadily from the beginning year to 2000 and increase rapidly from 2000 to 2018. The quadratic poisson model also covers most of the data with culvature but the upper 95% confidence interval at the start of the plot seems out of place. However, overall the model represents the data well.

Compute a confidence interval for the mean number of mass killing in 2019 using the linear and quadratic models. Assume that the populatino of the USA is 329,200,000 people. Comment briefly.

```
lin_poisson.fit <- glm(masskill ~ I(year - 1982), family = "poisson", offset
= log(popn/100), data = masskill_file)
new_data <-data.frame(year=2019, popn = 329.2)
lin_predict_2019 <- predict(lin_poisson.fit, new_data, se.fit = TRUE)
lin_predict_2019_high <- exp(lin_predict_2019$fit + lin_predict_2019$se.fit *
1.96)
lin_predict_2019_low <- exp(lin_predict_2019$fit - lin_predict_2019$se.fit *
1.96)
lin_predict_2019_high

##        1
## 10.97851

lin_predict_2019_low
```

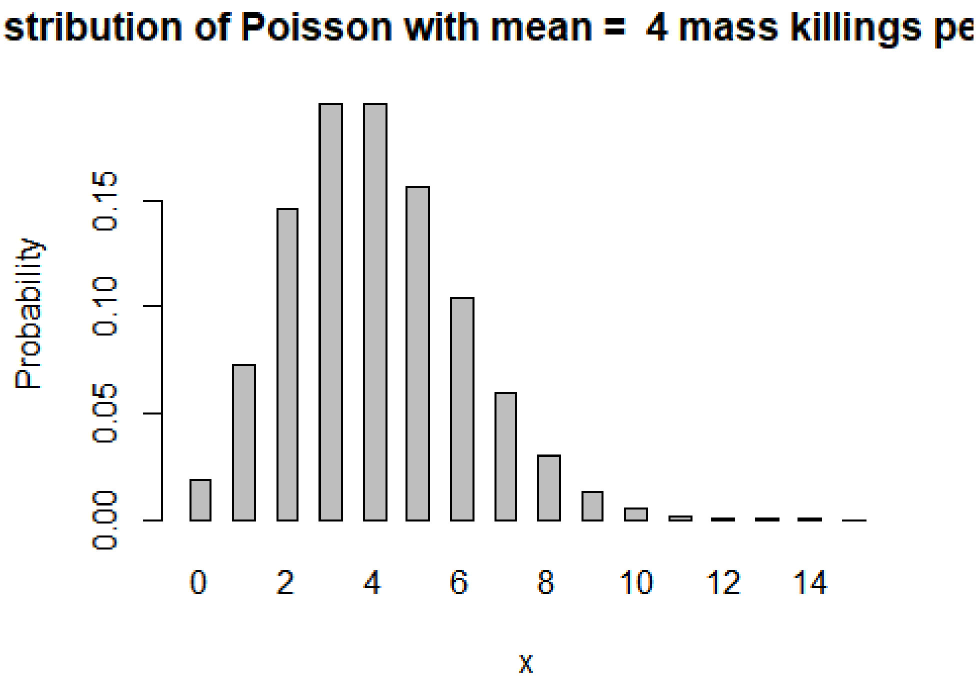
```
##        1
## 5.894407

exp(lin_predict_2019$fit)

##        1
## 8.044365

quad_predict_2019 <- predict(quad_poisson.fit, new_data, se.fit = TRUE)
quad_predict_2019_high <- exp(quad_predict_2019$fit +
quad_predict_2019$se.fit * 1.96)
quad_predict_2019_low <- exp(quad_predict_2019$fit - quad_predict_2019$se.fit
* 1.96)
quad_predict_2019_high

##        1
## 17.30245

quad_predict_2019_low

##        1
## 7.592244

exp(quad_predict_2019$fit)

##        1
## 11.46143
```

We estimate that the mean number of mass killing in 2019 using the linear poisson model with given USA 329.9 million population is somewhere between 6 people and 11 people. Whereae, we estimate that the mean number of mass killings in 2019 using the quadratic poisson model with given USA 329.9 million population is somewhere between 8 people and 17 people.

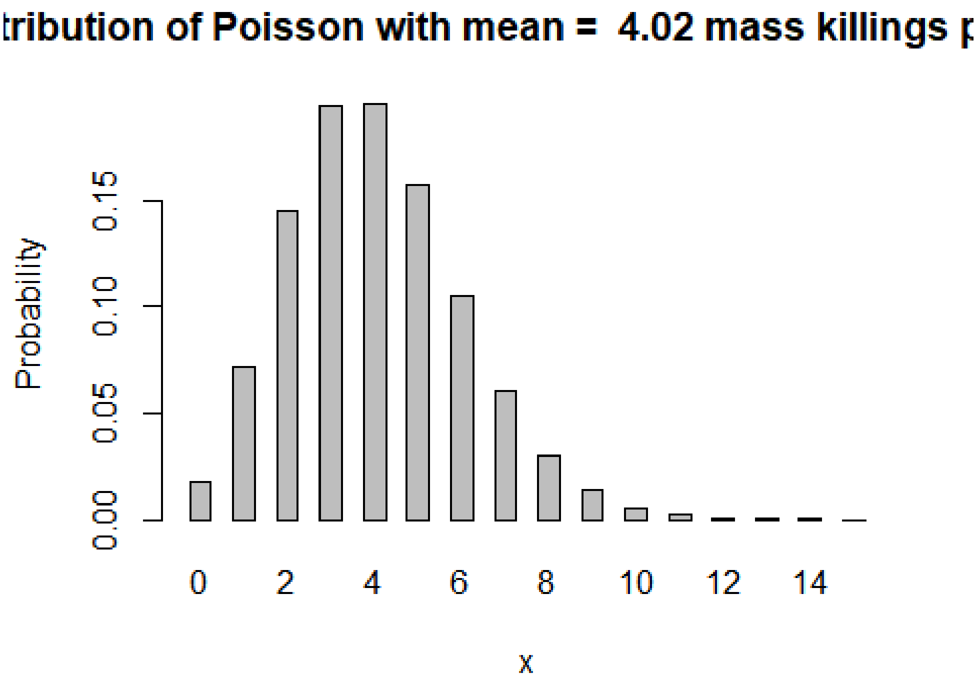
Use the linear and qudratic models predicted expected value and the code (changed by you), below, for the number of mass killings (adjusted for this half year scale) to see which of your models seems most appropriate for these data.

distrubution of Poisson (lambda = mean) distribution

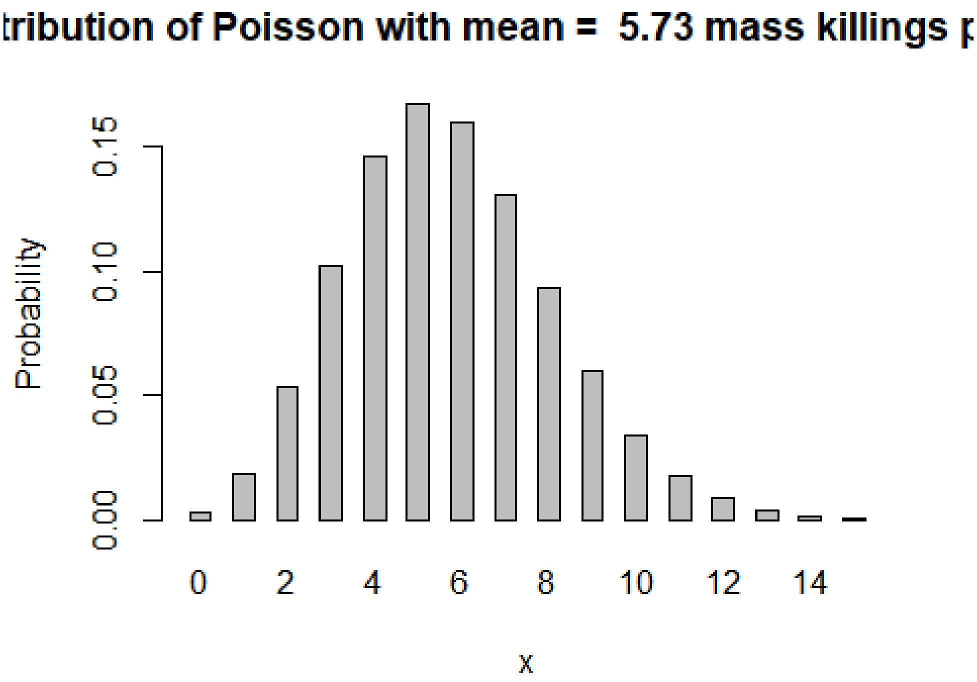
```
pred.mean = 4  #change this
barplot(dpois(0:15, pred.mean), ylab="Probability", xlab="x", space = 1,
names.arg = 0:15,
main=paste("Distribution of Poisson with mean = ",
          round(pred.mean,2), "mass killings per year"))
```



```
pred.mean = 8.044365/2 #change this
barplot(dpois(0:15, pred.mean), ylab="Probability", xlab="x", space = 1,
names.arg = 0:15,
main=paste("Distribution of Poisson with mean = ",
          round(pred.mean,2), "mass killings per year"))
```



```
pred.mean = 11.46143/2 #change this
barplot(dpois(0:15, pred.mean), ylab="Probability", xlab="x", space = 1,
names.arg = 0:15,
main=paste("Distribution of Poisson with mean = ",
          round(pred.mean,2), "mass killings per year"))
```



The prediction mean number of mass killings per year for linear poisson model is 4.0222 (4dp) whereas the prediction mean for quadratic poisson model is 5.7307 (4dp). Since the poisson distribution gives the probability of a mean number of events occuring in a fixed interval, we can safely assume that the rate of masskill per year will be the same as rate of masskill per half a year. We can see that both linear poisson model and the 2019 model are unimodal and right skewed. Whereas, the quadratic poisson model seems to be slightly skewed to the right and unimodal. The linear poisson model seems most appropriate for these data because it looks much similar to the actual mass killing incident and the rate of mass killings per half a year is alike with the real rate of mass killing per half a year.

As a consequence of the above analyses which of these two models ydo you believe is the best desription for these data. Comment, briefly, on what model you prefer and why?

With thorough analysis of the output of both models, I would prefer to use linear poisson model with offset population exposure over quadratic poisson model because the linear poisson model seems more suitable with the mass killing data. This is because the plot seems to be linearly increasing with fairly amount of scatter over the year between 1982 to 2015. It is only the recent year, 2017 and 2018, that there has been a high number masskill per 100 million people. Therefore, having a quadratic term would be excessive just for recent high mass killing data point. On top of that, from above analysis, we can see that the predicted linear poisson model was much closer to the model with the new data in terms of apperance of the model and mean number of mass killings per half a year. Furthermore, I belive it is also highly unlikely for the mass killing per 100 million people per year to increase rapidly so the poisson model wouldn’t be suitable to predict future mass killings per year.