Homework 2

Sarah Adilijiang

Problem 1

(a)

First, the paper analyzed the all-cause midlife (ages 45-54) mortality rates of several rich countries and found that there was a marked increase in the midlife mortality rate for white non-Hispanics in the United States between 1999 and 2013, which presented a reversal trend comparing with other rich countries and being unique to the United States.

Second, it analyzed the mortality rates from different causes and pointed out the three causes of death that accounted for the mortality reversal among white non-Hispanics, namely suicide, drug and alcohol poisoning, and chronic liver diseases and cirrhosis.

Also, the authors showed that people with less education have the most marked increases in mortality from suicide and poisonings.

Furthermore, it revealed that the rising midlife mortality rates of white non-Hispanics were paralleled by increases in midlife morbidity.

As a result, the authors stated that the above increased morbidity and mortality in midlife among white non-Hispanic were probably caused by growing distress in this population and the economic insecurity.

(b)

```
load("Mortality.Rdata")
str(Mortality)    # rate = Deaths/Population * 10^5
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    60 obs. of  9 variables:
##  $ Year      : Factor w/ 2 levels "1999","2013": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Ages      : num  45 45 45 46 46 46 47 47 47 48 ...
##  $ Race      : Factor w/ 2 levels "Black","White": 1 2 2 1 2 2 1 2 2 1 ...
##  $ Hisp      : Factor w/ 2 levels "YHisp","NHisp": 2 1 2 2 1 2 2 1 2 2 ...
##  $ Deaths    : num  2779 915 8304 2897 947 ...
##  $ Population: num  491705 354873 3166393 445148 329481 ...
##  $ rate      : num  565 258 262 651 287 ...
##  $ drug_alc  : int  268 119 930 235 151 894 217 117 840 205 ...
##  $ Suicide   : int  40 33 541 36 31 514 25 19 514 23 ...
```

```
# aggregate the dataset by Year & Ages
library(dplyr)
```

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

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

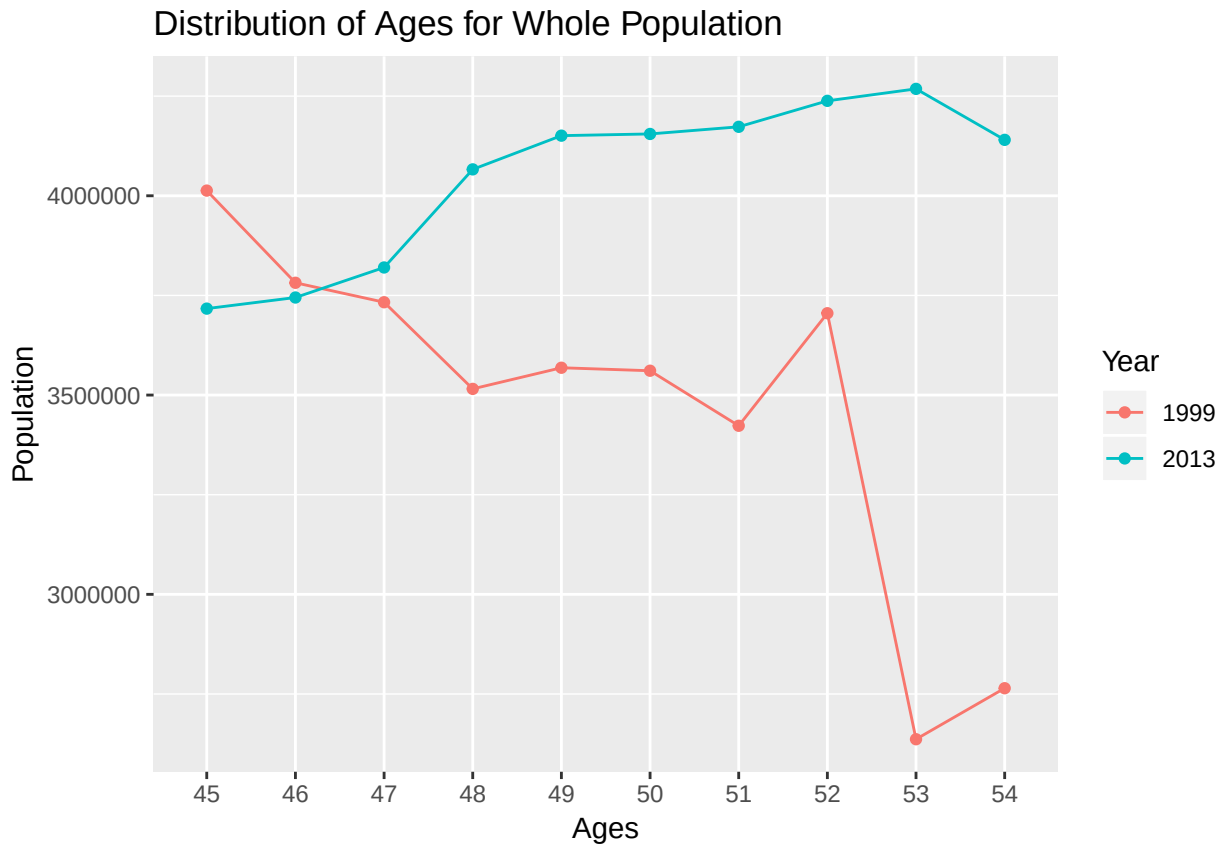
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
mortality_agg = Mortality %>% group_by(Year, Ages) %>% summarise(Population = sum(Population))
```

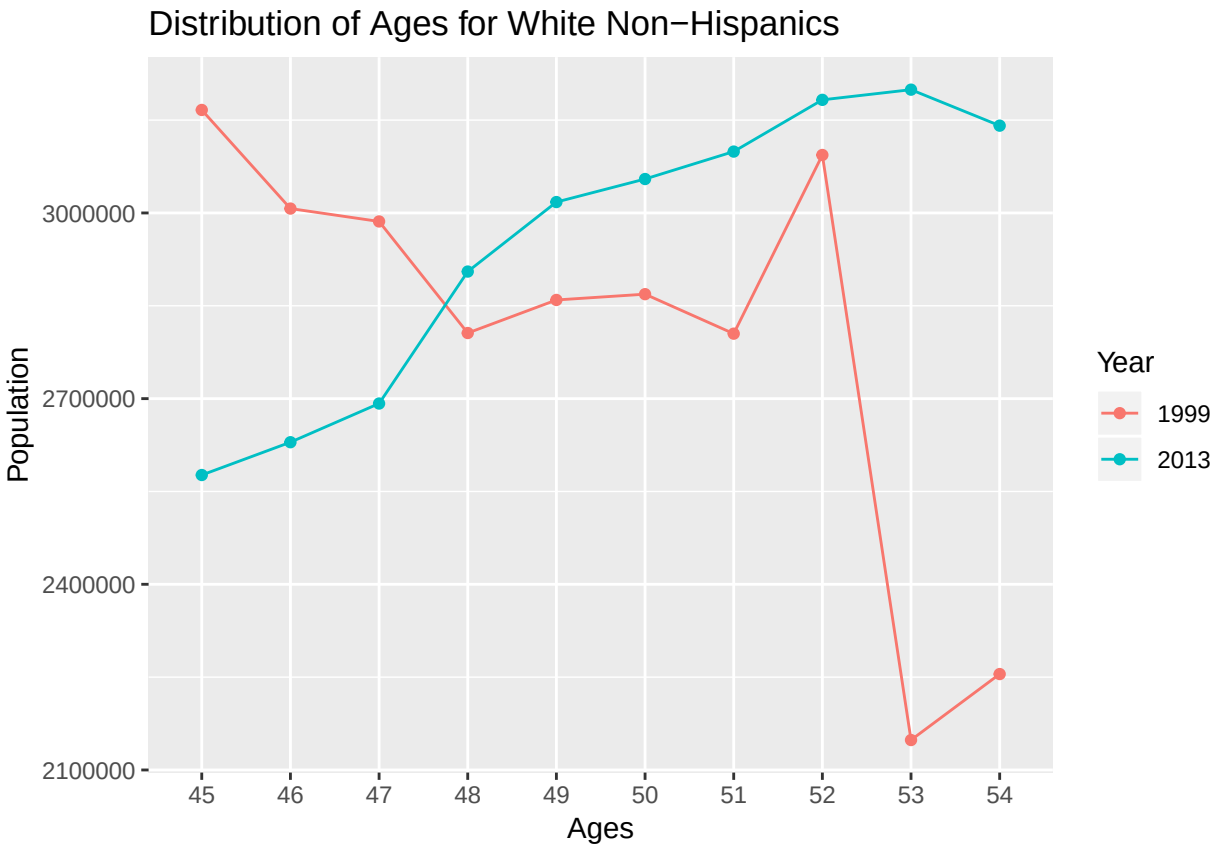
```r
# select the dataset for White Non-Hispanics
mortality_WNH = subset(Mortality, Race=="White" & Hisp=="NHisp")


# plot the distribution of "Ages" for whole population
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```r
ggplot(mortality_agg, aes(x=as.factor(Ages), y=Population, color=Year)) +
    geom_point() + geom_line(aes(group=Year)) +
    labs(x="Ages", title="Distribution of Ages for Whole Population")
```




```r
# plot the distribution of "Ages" for White Non-Hispanics
ggplot(mortality_WNH, aes(x=as.factor(Ages), y=Population, color=Year)) +
    geom_point() + geom_line(aes(group=Year)) +
    labs(x="Ages", title="Distribution of Ages for White Non-Hispanics")
```



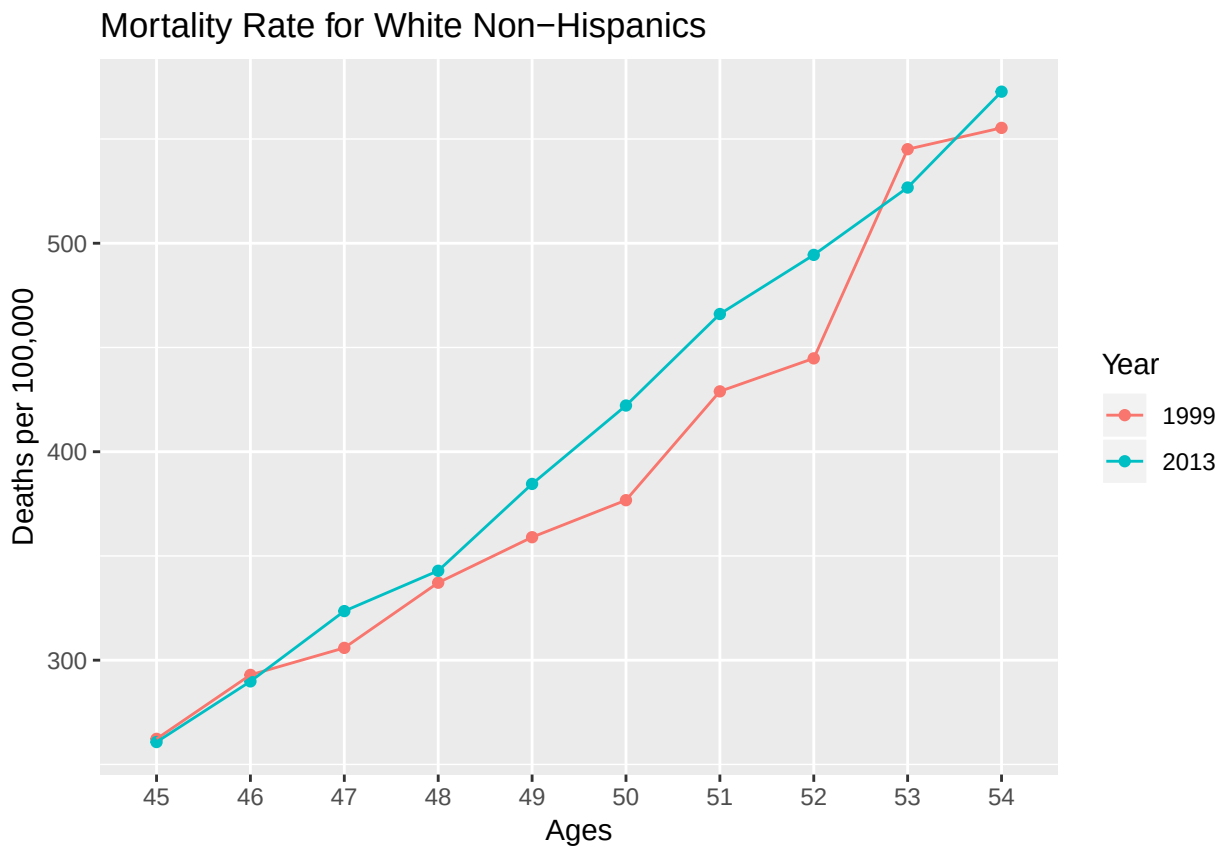
Answer:

For both the whole population and the white non-Hispanics subset, the distribution of ages in the 45 to 54 range shows different trends for the year 1999 and 2013. In the year 1999, the population has a generally decreasing trend as the age increases, while in the year 2013, the population has a generally increasing trend as the age increases.

This might affect the conclusion that mortality rates in the 45-54 age range increased among white non-hispanic Americans between 1999 and 2013, because there are more old people in the year 2013 than 1999, which might be one of the reasons for higher mortality rates in 2013 since elder people may have more risk of health problems.

(c)

```
# plot Motality rate ~ Ages, for both years
ggplot(mortality_WNH, aes(x=as.factor(Ages), y=rate, color=Year)) +
    geom_point() + geom_line(aes(group=Year)) +
    labs(x="Ages", y="Deaths per 100,000", title="Mortality Rate for White Non-Hispanics")
```



We can see from the plot that the mortality rate does increases as the ages increases both for 1999 and 2013. Since the year 2013 has more old people than 1999 as shown in question (a), we should add both the "Ages" and "Year" as covariates into the model to control for the age distribution.

Therefore, the generalized linear model (GLM) for binomial response here is:
likelihood: $P(Deaths_i = y_i|p_i) = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i-y_i}$, $y_i = 0, 1, ..., m_i$, where $Deaths_i$ is the number of deaths per 100,000 population, and m_i is the number of population in each group.
linear predictor: $\eta_i = \beta_0 + \beta_1 Ages_i + \beta_2 1_{Year_i=2013} + \epsilon_i$
link function (logit): $\eta_i = log\frac{p_i}{1-p_i}$, where mortality rate $p_i = y_i/m_i$
The binomial distribution is a case within the exponential distribution family.

```
# fit the new model adding "Ages" and interaction terms
model.glm = glm(cbind(Deaths, Population-Deaths)~Ages+Year, mortality_WNH, family=binomial)
summary(model.glm)
```

```
##
## Call:
## glm(formula = cbind(Deaths, Population - Deaths) ~ Ages + Year,
##     family = binomial, data = mortality_WNH)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9678  -1.7160   0.0805   1.2105   7.3466
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.7799441  0.0377151 -259.31   <2e-16 ***
```

```
## Ages         0.0849929  0.0007529  112.89   <2e-16 ***
## Year2013     0.0472687  0.0042082   11.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13571.97  on 19  degrees of freedom
## Residual deviance:   179.25  on 17  degrees of freedom
## AIC: 408.19
##
## Number of Fisher Scoring iterations: 3
```

```
# Chi-square test
anova(model.glm, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Deaths, Population - Deaths)
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     19    13572.0
## Ages  1  13266.4        18      305.5 < 2.2e-16 ***
## Year  1    126.3        17      179.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

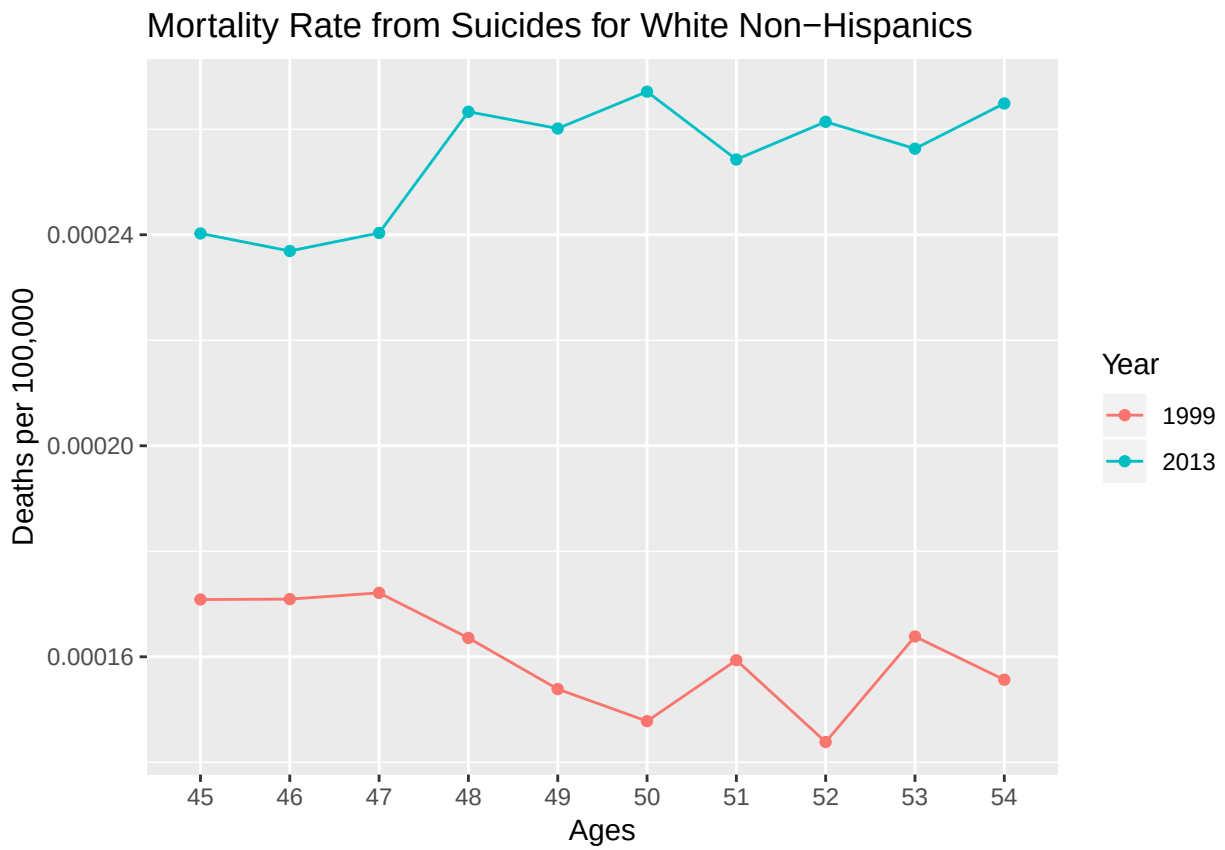
Answer:

In the summary, both the "Ages" and "Year2013" are significant predictors in this model. The difference in deviance chi-square test proves that the "Year" effect is highly significant.

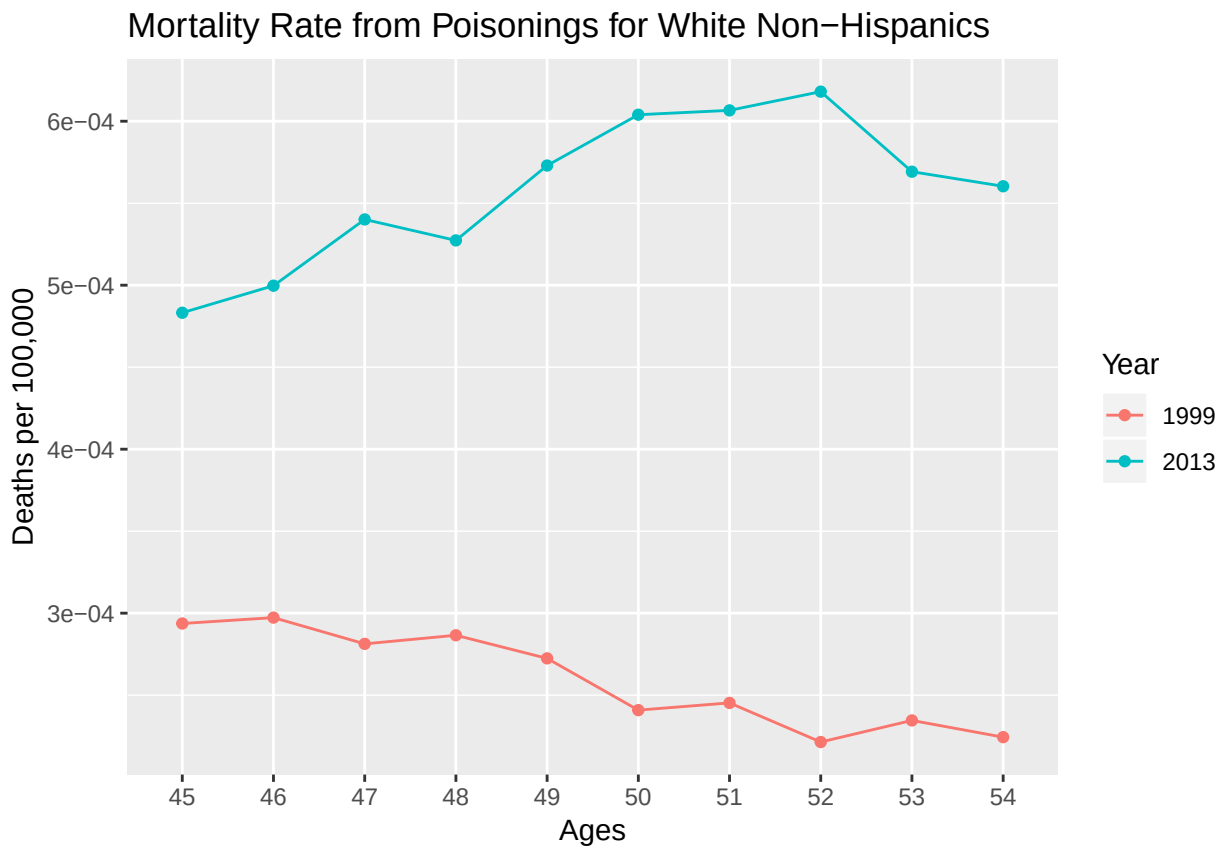
Based on this, the summary also shows that the "Year2013" is a significant positive predictor which means that the all-cause mortalily rate in 2013 is larger than that in 1999 when controlling the "Ages". Therefore, we can agree with the author's conclusion that the all-cause mortality rate of midlife white non-Hispanics has a marked increase between 1999 and 2013.

(d)

```
# plot Suicide mortality rate ~ Ages, for both years
ggplot(mortality_WNH, aes(x=as.factor(Ages), y=Suicide/Population, color=Year)) +
    geom_point() + geom_line(aes(group=Year)) +
    labs(x="Ages", y="Deaths per 100,000", title="Mortality Rate from Suicides for White Non-Hispanics"
```




```
# plot Poisoning mortality rate ~ Ages, for both years
ggplot(mortality_WNH, aes(x=as.factor(Ages), y=drug_alc/Population, color=Year)) +
    geom_point() + geom_line(aes(group=Year)) +
    labs(x="Ages", y="Deaths per 100,000", title="Mortality Rate from Poisonings for White Non-Hispanics
```



We can see from the plot that the mortality rate from suicides and poisonings both slightly decrease as the ages increases in the year 1999. However, the mortality rate from suicides and poisonings both slightly increase as the ages increases in the year 2013. Due to the two different trends, we should only only add both the "Ages" and "Year" as covariates into the model, but also consider the interaction terms of "Ages" and "Year".

Therefore, the generalized linear model (GLM) for binomial response here are:
For mortality rate from suicides:
likelihood: $P(Suicides_i = y_i|p_i) = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i-y_i},\ y_i = 0, 1, ..., m_i$, where $Suicides_i$ is the number of deaths per 100,000 population from suicides, and m_i is the number of population in each group.

For mortality rate from poisonings:
likelihood: $P(Poisonings_i = y_i|p_i) = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i-y_i},\ y_i = 0, 1, ..., m_i$, where $Suicides_i$ is the number of deaths per 100,000 population from poisonings, and m_i is the number of population in each group.

But the forms of linear predictors and link functions are the same:
linear predictor: $\eta_i = \beta_0 + \beta_1 Ages_i + \beta_2 1_{Year_i=2013} + \beta_3 Ages_i \times 1_{Year_i=2013} + \epsilon_i$
link function (logit): $\eta_i = log\frac{p_i}{1-p_i}$, where mortality rate $p_i = y_i/m_i$

```
# fit a model for "Suicide"
model.glm_su = glm(cbind(Suicide, Population-Suicide)~Ages*Year, mortality_WNH, family=binomial)
summary(model.glm_su)
```

```
##
## Call:
## glm(formula = cbind(Suicide, Population - Suicide) ~ Ages * Year,
##     family = binomial, data = mortality_WNH)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.44235  -0.81818   0.04646   0.75372   1.44309
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.015422   0.260706 -30.745  < 2e-16 ***
## Ages          -0.014697   0.005297  -2.775 0.005526 **
## Year2013      -0.760498   0.330319  -2.302 0.021318 *
## Ages:Year2013  0.024800   0.006678   3.713 0.000204 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 656.145  on 19  degrees of freedom
## Residual deviance:  15.154  on 16  degrees of freedom
## AIC: 187.05
##
## Number of Fisher Scoring iterations: 3
```

```
# Chi-square test
anova(model.glm_su, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Suicide, Population - Suicide)
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         19     656.14
## Ages       1     5.06        18     651.09 0.0245463 *
## Year       1   622.13        17      28.96 < 2.2e-16 ***
## Ages:Year  1    13.81        16      15.15 0.0002024 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# difference between 1999 and 2013 at different ages
beta = coef(model.glm_su)
beta[3] + beta[4]* mortality_WNH$Ages
```

```
##  [1] 0.3554993 0.3802992 0.4050992 0.4298991 0.4546991 0.4794990 0.5042990
##  [8] 0.5290989 0.5538988 0.5786988 0.3554993 0.3802992 0.4050992 0.4298991
## [15] 0.4546991 0.4794990 0.5042990 0.5290989 0.5538988 0.5786988
```

Answer:

In the summary, all the predictors are significant. The difference in deviance chi-square test proves that the "Year" effect and the interaction term "Ages:Year" are both highly significant.

Based on this, the summary also shows that when controlling the "Year", in 1999, "Ages" has a slope of $\hat{\beta}_1$ = -0.014697, indicating a negative effect of "Ages" on the mortalitly rate from suicides. In contrast, in 2013,

"Ages" has a slope of $\hat{\beta}_1 + \hat{\beta}_3$ = -0.014697 + 0.024800 = 0.010103, indicating a positive effect of "Ages" on the mortalitly rate from suicides. These are consistent with the discovers in the previous plots.

On the other hand, when controlling the "Ages", the mortality rate from suicides in 2013 is $\hat{\beta}_2 + \hat{\beta}_3 Ages_i$ larger than that in 1999. The values of $\hat{\beta}_2 + \hat{\beta}_3 Ages_i$ shows that though these differences between 1999 and 2013 are different at each age, they are all positive, which means the mortality rate from suicides in 2013 is larger than that of 1999 at all ages.

```
# fit a model for "drug_alc"
model.glm_dr = glm(cbind(drug_alc, Population-drug_alc)~Ages*Year, mortality_WNH, family=binomial)
summary(model.glm_dr)
```

```
##
## Call:
## glm(formula = cbind(drug_alc, Population - drug_alc) ~ Ages *
##     Year, family = binomial, data = mortality_WNH)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4892  -1.3919   0.2928   1.1284   2.9609
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.479704   0.204377 -31.705  < 2e-16 ***
## Ages          -0.036021   0.004167  -8.645  < 2e-16 ***
## Year2013      -1.961119   0.246200  -7.966 1.64e-15 ***
## Ages:Year2013  0.055215   0.004992  11.061  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3376.59  on 19  degrees of freedom
## Residual deviance:   56.36  on 16  degrees of freedom
## AIC: 240.88
##
## Number of Fisher Scoring iterations: 3
```

```
# Chi-square test
anova(model.glm_dr, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(drug_alc, Population - drug_alc)
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         19     3376.6
## Ages       1     29.8        18     3346.8 4.856e-08 ***
## Year       1   3167.5        17      179.3 < 2.2e-16 ***
## Ages:Year  1    122.9        16       56.4 < 2.2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# difference between 1999 and 2013 at different ages
beta = coef(model.glm_dr)
beta[3] + beta[4]* mortality_WNH$Ages
```

```
##  [1] 0.5235442 0.5787590 0.6339737 0.6891884 0.7444032 0.7996179 0.8548327
##  [8] 0.9100474 0.9652621 1.0204769 0.5235442 0.5787590 0.6339737 0.6891884
## [15] 0.7444032 0.7996179 0.8548327 0.9100474 0.9652621 1.0204769
```

Answer:

In the summary, all the predictors are significant. The difference in deviance chi-square test proves that the "Year" effect and the interaction term "Ages:Year" are both highly significant.

Based on this, the summary also shows that when controlling the "Year", in 1999, "Ages" has a slope of $\hat{\beta}_1$ = -0.036021, indicating a negative effect of "Ages" on the mortalitly rate from poisonings. In contrast, in 2013, "Ages" has a slope of $\hat{\beta}_1 + \hat{\beta}_3$ = -0.036021 + 0.055215 = 0.019194, indicating a positive effect of "Ages" on the mortalitly rate from poisonings. These are consistent with the discovers in the previous plots.

On the other hand, when controlling the "Ages", the mortality rate from suicides in 2013 is $\hat{\beta}_2 + \hat{\beta}_3 Ages_i$ larger than that in 1999. The values of $\hat{\beta}_2 + \hat{\beta}_3 Ages_i$ shows that though these differences between 1999 and 2013 are different at each age, they are all positive, which means the mortality rate from poisonings in 2013 is larger than that of 1999 at all ages.