**STAT34700-HW2**

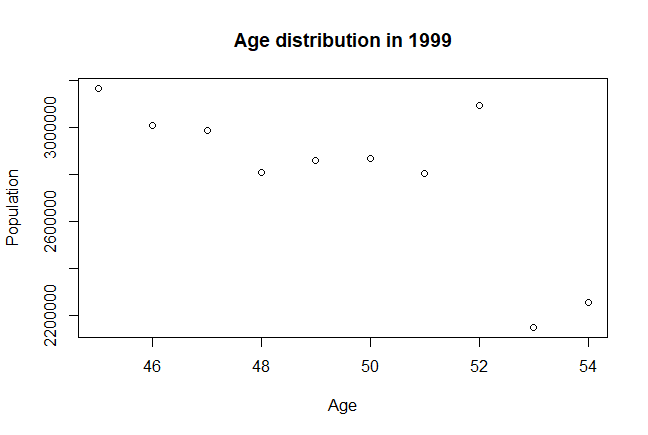
**Problem 1**

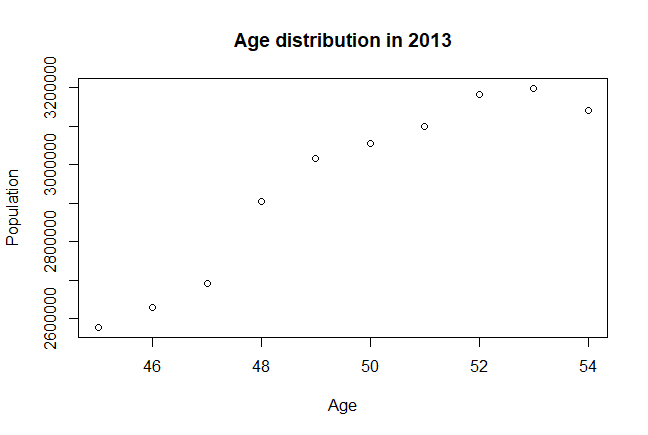
**(a)**

In this paper, it concludes that there is a marked increase in the all-cause mortality of middle-age (age 45 to 54) white non-Hispanic men and women in the United States between 1999 and 2013. This obvious increase of mortality reversed decades of progress of mortality and was unique in U.S. This midlife mortality increases only happened among white non-Hispanics so in other groups we continued to see mortality rates fall. It was result from increasing death rates from drug and alcohol poisonings, suicide, and chronic liver diseases and cirrhosis. For different education levels, those with less education saw the most marked increases. The increases in midlife morbidity parallel the rising midlife mortality rates of white non-Hispanics. In addition, there are some potential economic causes and consequences of this deterioration.

**(b)**

We plot the age distributions in 1999 and 2013:

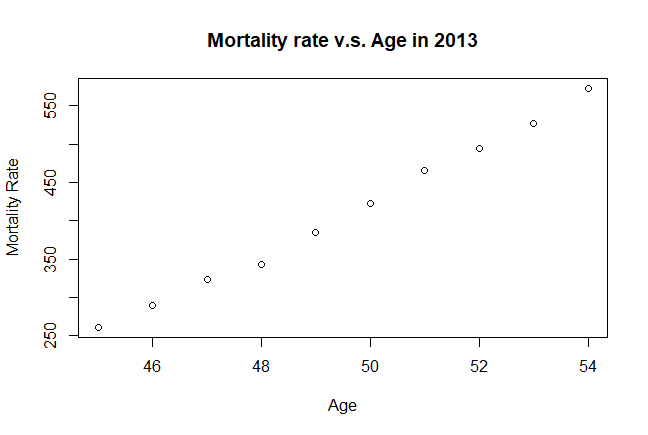




From the two plots we can see that in 1999, the population of people with age above 52 is very small, compared to other age groups. While in 2013, the population of people with age above 52 is large, compared to other age groups.

We also plot the mortality rate against age in 1999 and 2013:





From these two plots we can see that in both 1999 and 2013, the mortality rate increases as age is increasing. In other words, for people with age above 52, their mortality rate is very large compared to other age group in both 1999 and 2013.

**Conclusion:**

Form the plots of age distribution in 1999 and 2013, the population of people with age above 52 is very small in 1999 but very large in 2013. Therefore, there will be less mortality for population with age 45-54 in 1999 than in 2013. However, for each smaller age group, the mortality may not increase from 1999 to 2013. As a result, we can say that the increase of mortality rate between 1999 and 2013 may also due to the different age distributions of midlives in these years but not only those reasons listed in the paper.

**(c)**

To control the age distribution using a generalized linear model, we can fit a binomial GLM with *Deaths* as the binomial response, *Year* as a categorical predictor, and *Ages* as a numerical predictor. In this generalized linear model, we use logit link function and binomial distribution family. We can test the significance of predictor *Year* to determine whether there is a marked increase of midlife mortality between 1999 and 2013. In addition, we can test the significance of predictor *Ages* to determine whether the age distribution can affect the midlife mortality.

**The GLM formula:**

We fit the model to data of white non-Hispanic:

fit\_1c = glm(cbind(Deaths,Alive) ~ Ages + Year, family=binomial, data=mort)

**Output:**

> sumary(fit\_1c)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -9.77994415 0.03771514 -259.311 < 2.2e-16

Ages 0.08499287 0.00075291 112.886 < 2.2e-16

Year2013 0.04726874 0.00420821 11.232 < 2.2e-16

n = 20 p = 3

Deviance = 179.25177 Null Deviance = 13571.97185 (Difference = 13392.72008)

From the output we can see that the p-values of coefficients corresponding to *Ages* and *Year* are both small. As a result, the hypothesis that *Ages* and *Year* are not significant can be rejected. Then, we can conclude that both *Age* and *Year* are significant in this binomial GLM.

**Conclusion:**

From the conclusion we got above, we can say that there is a marked increase of midlife mortality between 1999 and 2013 and the age distribution can also affect the midlife mortality.

Therefore, we can partially agree with the conclusion of the paper: the midlife mortality for white non-Hispanic increased from 1999 to 2013. However, the age distribution is also a reason which can explain this midlife mortality increase.

**(d)**

**1) Analysis deaths from suicides**

We fit a binomial GLM with *Suicide (number of death from suicide)* as the binomial response, *Year* as a categorical predictor, and *Ages* as a numerical predictor. In this generalized linear model, we use logit link function and binomial distribution family. We can test the significance of predictor *Year* to determine whether there is a marked increase of midlife mortality from suicide between 1999 and 2013. In addition, we can test the significance of predictor *Ages* to determine whether the age distribution can affect the midlife mortality from suicide.

We fit the model to data of white non-Hispanic:

fit\_1ds = glm(cbind(Suicide,NotSuicide) ~ Ages + Year, family=binomial, data=mort)

**The GLM formula:**

**Output:**

> sumary(fit\_1ds)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.78275064 0.15935618 -55.1140 <2e-16

Ages 0.00089902 0.00322058 0.2791 0.7801

Year2013 0.46473558 0.01891800 24.5658 <2e-16

n = 20 p = 3

Deviance = 28.96261 Null Deviance = 656.14471 (Difference = 627.18211)

> mort$NotDrug = mort$Population - mort$drug\_alc

> fit\_1dd = glm(cbind(drug\_alc,NotDrug) ~ Ages + Year, family=binomial, data=mort)

From the output we can see that the p-value of coefficients corresponding to *Year* are smaller

than 0.05 but the p-value of coefficients corresponding to *Ages* are larger than 0.05. As a result,

the null hypothesis that *Year* is not significant can be rejected, and the null hypothesis that *Ages*   
is not significant cannot be rejected at significance level 0.05. Then, we can conclude that *Year*

is significant but *Ages* is not significant in this binomial GLM.

**Conclusion:**

From the conclusion we got above, we can say that there is a marked increase of midlife mortality from suicide between 1999 and 2013 and the age distribution cannot affect the midlife mortality from suicide. Therefore, we can partially agree with the conclusion of the paper: the midlife mortality from suicide for white non-Hispanic increased from 1999 to 2013. However, the age distribution is not a main reason which can explain this midlife mortality increase from suicide.

**2) Analysis deaths from substance abuse**

We fit a binomial GLM with *drug\_alc (number of deaths from substance abuse* as the binomial response, *Year* as a categorical predictor, and *Ages* as a numerical predictor. In this generalized linear model, we use logit link function and binomial distribution family. We can test the significance of predictor *Year* to determine whether there is a marked increase of midlife mortality from substance abuse between 1999 and 2013. In addition, we can test the significance of predictor *Ages* to determine whether the age distribution can affect the midlife mortality from substance abuse.

We fit the model to data of white non-Hispanic:

fit\_1dd = glm(cbind(drug\_alc,NotDrug) ~ Ages + Year, family=binomial, data=mort)

**The GLM formula:**

**Output:**

> sumary(fit\_1dd)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.3667053 0.1131836 -73.9215 <2e-16

Ages 0.0023932 0.0022847 1.0475 0.2949

Year2013 0.7624134 0.0140767 54.1613 <2e-16

n = 20 p = 3

Deviance = 179.29628 Null Deviance = 3376.59312 (Difference = 3197.29684)

From the output we can see that the p-value of coefficients corresponding to *Year* are smaller

than 0.05 but the p-value of coefficients corresponding to *Ages* are larger than 0.05. As a result,

the null hypothesis that *Year* is not significant can be rejected, and the null hypothesis that *Age*

is not significant cannot be rejected at significance level 0.05. Then, we can conclude that *Year* issignificant but *Ages* is not significant in this binomial GLM.

**Conclusion:**

From the conclusion we got above, we can say that there is a marked increase of midlife mortality from substance abuse between 1999 and 2013 and the age distribution cannot affect the midlife mortality from substance abuse. Therefore, we can partially agree with the conclusion of the paper: the midlife mortality from substance abuse for white non-Hispanic increased from 1999 to 2013. However, the age distribution is not a main reason which can explain this midlife mortality increase from substance abuse.

**R Code**

# Problem 1 ---------------------------------------------------------------

load("Mortality.RData")

install.packages("faraway")

library(faraway)

#b

data1999=subset(Mortality,Year==1999)

data2013=subset(Mortality,Year==2013)

#Plot the age distribution in 1999 and 2013

popu1999=NULL

for (i in 1:30){

if (i %% 3 == 0) {

popu1999 = c(popu1999,data1999$Population[i])

}

}

age=c(45,46,47,48,49,50,51,52,53,54)

plot(age,popu1999,xlab="Age",ylab="Population",main="Age distribution in 1999")

popu2013=NULL

for (i in 1:30){

if (i %% 3 ==0){

popu2013 = c(popu2013,data2013$Population[i])

}

}

age=c(45,46,47,48,49,50,51,52,53,54)

plot(age,popu2013,xlab="Age",ylab="Population",main="Age distribution in 2013")

#Plot the mortality rate against age in 1999 and 2013

mort1999=NULL

for (i in 1:30){

if (i %% 3 == 0) {

mort1999 = c(mort1999,data1999$rate[i])

}

}

age=c(45,46,47,48,49,50,51,52,53,54)

plot(age,mort1999,xlab="Age",ylab="Mortality Rate",main="Mortality rate v.s. Age in 1999")

mort2013=NULL

for (i in 1:30){

if (i %% 3 == 0) {

mort2013 = c(mort2013,data2013$rate[i])

}

}

age=c(45,46,47,48,49,50,51,52,53,54)

plot(age,mort2013,xlab="Age",ylab="Mortality Rate",main="Mortality rate v.s. Age in 2013")

#c

#select data of white non-Hispanic

a = subset(Mortality, Race == "White")

mort = subset(a, Hisp == "NHisp")

mort$Alive = mort$Population - mort$Deaths

#fit the binomial model

fit\_1c = glm(cbind(Deaths,Alive) ~ Ages + Year, family=binomial, data=mort)

sumary(fit\_1c)

#d

#analysis for suicide

mort$NotSuicide = mort$Population - mort$Suicide

fit\_1ds = glm(cbind(Suicide,NotSuicide) ~ Ages + Year, family=binomial, data=mort)

sumary(fit\_1ds)

#analysis for substances abuse

mort$NotDrug = mort$Population - mort$drug\_alc

fit\_1dd = glm(cbind(drug\_alc,NotDrug) ~ Ages + Year, family=binomial, data=mort)

sumary(fit\_1dd)