Homework 7: Poisson Regression

公衛二 梁嫚芳 B07801003

Based on the count data to answer the following questions.

1. Plot the count data, where the x-axis represents Age and the y-axis represents Deaths *per 100000 person-year*. Do you think Smoke is a risk factor of Deaths? Why?

|  |
| --- |
| 圖中 smokers和non-smokers的死亡率皆隨著Age的增加而上升，除了Age為80歲的組別以外。每組smokers的死亡率皆較non-smokers高，二者間的差距隨著年齡增加而有變大的趨勢。因此我認為Smoke是Deaths的風險因子。 |

1. Fit a Poisson regression model to answer if Smoke is a risk factor of Deaths or not. Write down the model and the corresponding null hypothesis to make conclusion.

|  |
| --- |
| Y: Deaths  X1: Age  X2: Smoke={0:smoke, 1:non-smoke}  Y|X~Poisson(m λx) with m λx = E[Y|X]  ln(λx) = β0+β1X1+β2X2 → λx = e β0+β1X1+β2X2  β0 : 模型的截距  β1 : 同Smoke下，Age每增加一單位，risk ratio of death平均增加exp(β1)倍。  β2 : 同Age下，有Smoke較無Smoke對risk ratio of death平均增加exp(β2)倍。 |
| H0: β2=0 H1: β2≠0  p-value=0.00015 < α(0.05)，拒絕H0，β2與0達統計顯著的差異，Smoke會影響Death。 |

1. The plot in Problem-1 suggests the **existence of interaction between Age and Smoke**. Fit a Poisson regression model to answer if Deaths is affected by Smoke or not. Write down the model and the corresponding null hypothesis. Calculate the *LRT statistic* to make conclusion.

|  |
| --- |
| Y: Deaths  X1: Age  X2: Smoke={0,1}  Y|X~Poisson(m λx) with m λx = E[Y|X]  Full model (md1) : ln(λx) = β0+β1X1+β2X2+β3(X1 X2)  Reduced model (md2) : ln(λx) = β0+β1X1  β0 : 模型的截距  β1 : 同Smoke下，Age每增加一單位，risk ratio of death平均增加exp(β1)倍。  β2 : 同Age下，有Smoke較無Smoke對risk ratio of death平均增加exp(β2)倍。  β3 : Smoke和Age交互作用下對Deaths的影響 |
| H0: β2=β3=0 H1: β2, β3不均為0  LRT *statistic* : D=25.118 > Χ2=5.991465  p-value=3.513e-06 < α(0.05)，拒絕H0，β1, β3達統計顯著地不均為0，Smoke會影響Death。 |

1. Plot the two estimated (*per 100000 person-year*) mean responses of Smokers and Non-smokers at different Ages under the model of Problem-3. Comparing the plot with Problem-1, do you think the Poisson regression model is adequately specified? Why? 這題被扣兩分

|  |
| --- |
| 右圖將mean responses與plot of Problem-1疊圖，可看出mean response 大致能預測不同Age下，smoke與non-smoke的Deaths，且Promblem-3的圖比起Problem-1的圖更能解釋Deaths、smoke與age的關係。因此我認為Poisson regression model is adequately specified. |

##1

code

#build data.frame of smokers and nonsmokers

smokers <- data.frame(Age=c(40,50,60,70,80),Deaths=c(32,104,206,186,102),py=c(52407,43248,28612,12663,5317),smoke=c(1,1,1,1,1))

nonsmokers<-data.frame(Age=c(40,50,60,70,80),Deaths=c(2,12,28,28,31),py=c(18790,10673,5710,2585,1462),smoke=c(0,0,0,0,0))

#create new variable of unit person-year

smokers$Deaths\_py<-smokers$Deaths/smokers$py\*100000

nonsmokers$Deaths\_py<-nonsmokers$Deaths/nonsmokers$py\*100000

#plot the data point

plot(smokers$Age,smokers$Deaths\_py,pch=1,xlim=c(40,85),ylim=c(0,4000),xlab="Age",ylab="Deaths per 100000 person-year")

points(nonsmokers$Age,nonsmokerss$Deaths\_py,xlim=c(40,85),ylim=c(0,4000),pch=16)

legend("bottomright",legend=c("smokers","non-smokers"),pch=c(1,16))

##2

#fit the poisson model

countdata<-rbind(smokers,nonsmokers)

fit<-glm(Deaths~smoke+Age,offset=log(py),family=poisson,data=countdata)

summary(fit)

##3

#testing

md1<-glm(Deaths~smoke\*Age,offset=log(py),family=poisson,data=countdata)

md2<-glm(Deaths~Age,offset=log(py),family=poisson,data=countdata)

anova(md2,md1,test="Chisq")

qchisq(0.95,df=2)

##4

#plot mean response

curve(predict(md1,data.frame(Age=x,smoke=1,py=100000),type="response"),lty=1,xlim=c(40,85),ylim=c(0,4000),xlab="Age",ylab="Mean response")

curve(predict(md1,data.frame(Age=x,smoke=0,py=100000),type="response"),lty=2,xlim=c(40,85),ylim=c(0,4000),xlab="Age",ylab="Mean response",add=T)

legend("bottomright",legend=c("smokers","non-smokers"),lty=c(1,2))

#plots overlay

plot(smokers$Age,smokers$Deaths\_py,pch=1,xlim=c(40,85),ylim=c(0,4000),xlab="Age",ylab="Deaths per 100000 person-year")

points(nonsmokers$Age,nonsmokers$Deaths\_py,xlim=c(40,85),ylim=c(0,4000),pch=16)

curve(predict(md1,data.frame(Age=x,smoke=1,py=100000),type="response"),lty=1,xlim=c(40,85),ylim=c(0,4000),xlab="Age",ylab="Mean response",add=T)

curve(predict(md1,data.frame(Age=x,smoke=0,py=100000),type="response"),lty=2,xlim=c(40,85),ylim=c(0,4000),xlab="Age",ylab="Mean response",add=T)

legend("bottomright",legend=c("smokers","non-smokers"),lty=c(1,2))

legend("topright",legend=c("smokers","non-smokers"),pch=c(1,16))