**實習課作業4**

**Deadline: 2019/12/19(四) 14:00**

|  |
| --- |
| 1. 請使用檔案cholera.csv 完成以下作業 2. 將 R code 及重點 output 貼到 word 檔上，寫上適當敘述後上傳至ceiba 作業區。 |

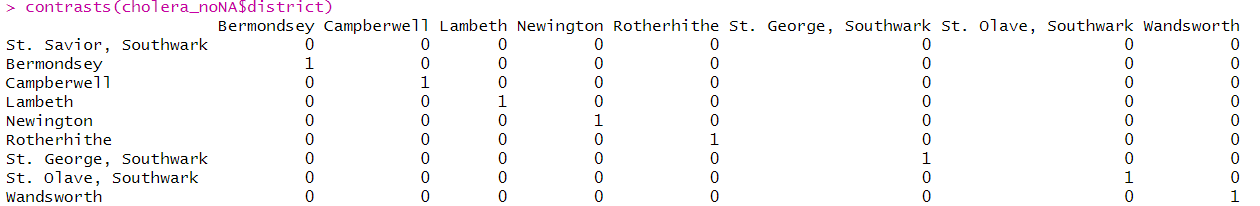
配分：一題50分

為探討19世紀倫敦霍亂是否與地區及自來水公司有關，請根據以上敘述寫出 model 的樣子，須定義符號。

|  |
| --- |
| 資料整理：   1. 將原本的data 匯入並命名為cholera 2. 建立新data.frame將之命名為cholera\_new 3. 將cholera中同 district與同company的DeathIndicator=1的Count數量匯入為cholera\_new$deaths(同地區下，同自來水公司的死亡人數) 4. 將cholera中同 district與同company的Count數量加總匯入為cholera\_new$py(同地區下，同自來水公司的總人數) 5. 將cholera\_new$py為0的資料刪除並命名為cholera\_noNA，以fit model |
| Code |
| ##1  cholera<-read.csv(file.choose())  # build new data.frame  district\_levels<-c(levels(cholera$District))  company\_levels<-c(1,2)  district<-vector()  company<-vector()  deaths<-vector()  py<-vector()  k<-rep(c(1:9),each=2)  company<-rep(c(1:2),9)  for(i in c(1:18)){  district[i]<-district\_levels[k[i]]  }  for(i in seq(1,18,2)){  deaths[i]<-subset(cholera,cholera$District==district\_levels[k[i]] & cholera$Company==company\_levels[1] & cholera$DeathIndicator==1)$ Count  py[i]<-sum(subset(cholera,cholera$District==district\_levels[k[i]] & cholera$Company==company\_levels[1])$ Count)  }  for(i in seq(2,18,2)){  deaths[i]<-subset(cholera,cholera$District==district\_levels[k[i]] & cholera$Company==company\_levels[2] & cholera$DeathIndicator==1)$ Count  py[i]<-sum(subset(cholera,cholera$District==district\_levels[k[i]] & cholera$Company==company\_levels[2])$ Count)  }  cholera\_new<-data.frame(district,company,deaths,py)  cholera\_new$district<-as.factor(district)  cholera\_new$company<-as.factor(company)  # delete which total counts=0  cholera\_noNA<-subset(cholera\_new,cholera\_new$py!=0) |

|  |
| --- |
| Fit the model |
| Output |
|  |
| code |
| cholera\_noNA$district<-relevel(cholera\_noNA$district,"St. Savior, Southwark")  contrasts(cholera\_noNA$district)  contrasts(cholera\_noNA$company)  fit<-glm(deaths~company+district,offset=log(py),family=poisson,data=cholera\_noNA)  summary(fit) |

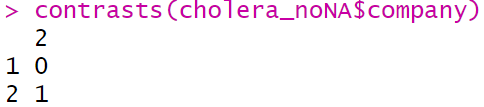
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| District | X1(1) | X1(2) | X1(3) | X1(4) | X1(5) | X1(6) | X1(7) | X1(8) |



Y|X~Poisson(m·λx) with m·λx = E[Y|X]

ln(λx)=β0+β1(1)X1(1)+β1(2)X1(2)+β1(3)X1(3)+β1(4)X1(4)+ β1(5)X1(5)+β1(6)X1(6)+β1(7)X1(7)+β1(8)X1(8)+ β2X2

Y : Deaths

X1 : District

X2 : Company={0:company1,1:company2}

β0 : 模型的截距

β1(1) : 相同公司下，Bermondsey較St. Savior, Southwark的deaths rate多e β1(1)倍

β1(2) : 相同公司下，Campberwell較St. Savior, Southwark的deaths rate多e β1(2)倍

β1(3) : 相同公司下，Lambeth較St. Savior, Southwark的deaths rate多e β1(3)倍

β1(4) : 相同公司下，Newington較St. Savior, Southwark的deaths rate多e β1(4)倍

β1(5) : 相同公司下，Rotherhithe較St. Savior, Southwark的deaths rate多e β1(5)倍

β1(6) : 相同公司下，St. George, Southwark較St. Savior, Southwark的deaths rate多e β1(6)倍

β1(7) : 相同公司下，St. Olave, Southwark較St. Savior, Southwark的deaths rate多e β1(7)倍

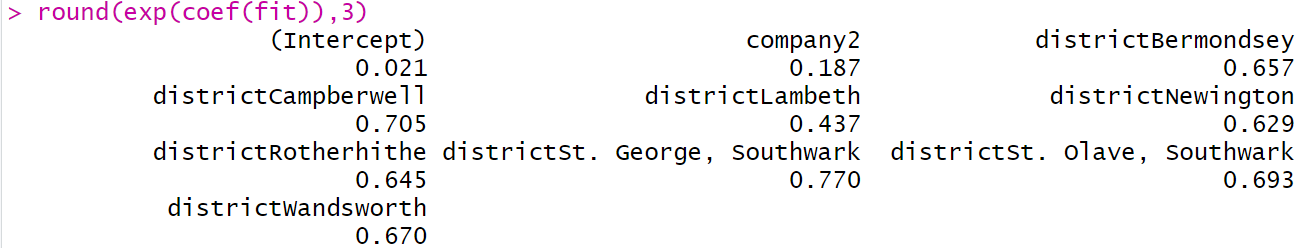
β1(8) : 相同公司下，Wandsworth較St. Savior, Southwark的deaths rate多e β1(8)倍

β2 : 相同地區下，company2較company1的deaths rate多e β2倍

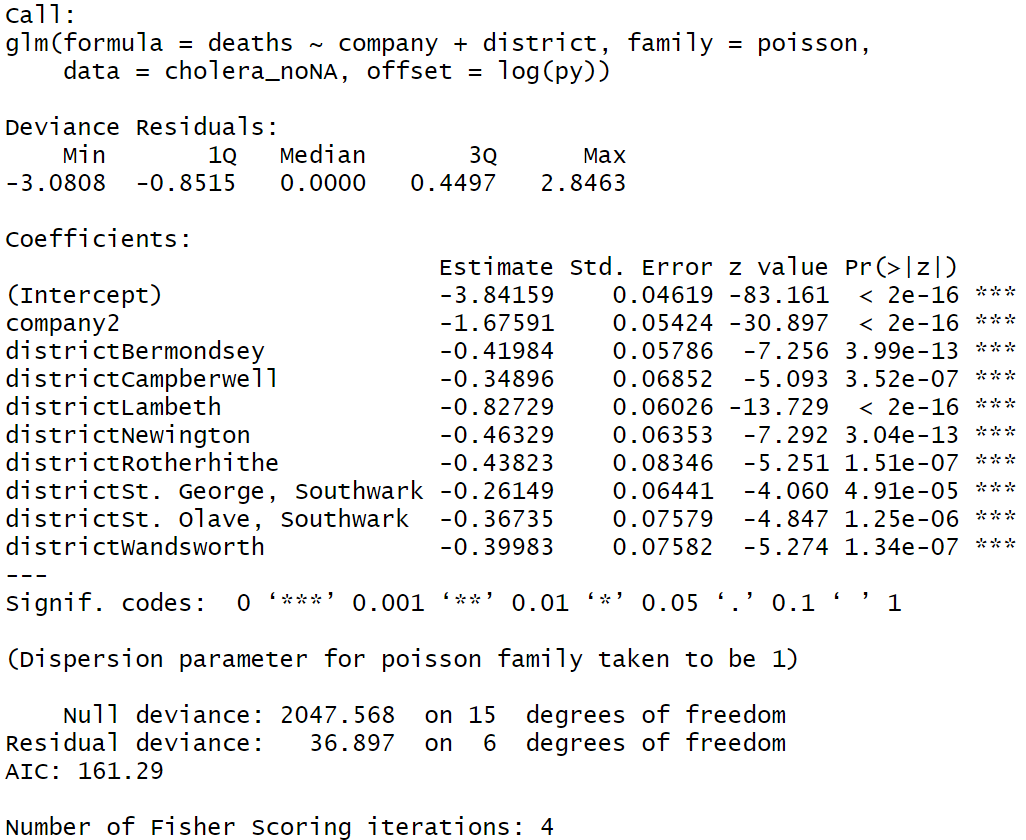
1. 在調整自來水公司後，地區對霍亂死亡率的影響為何？

Code:

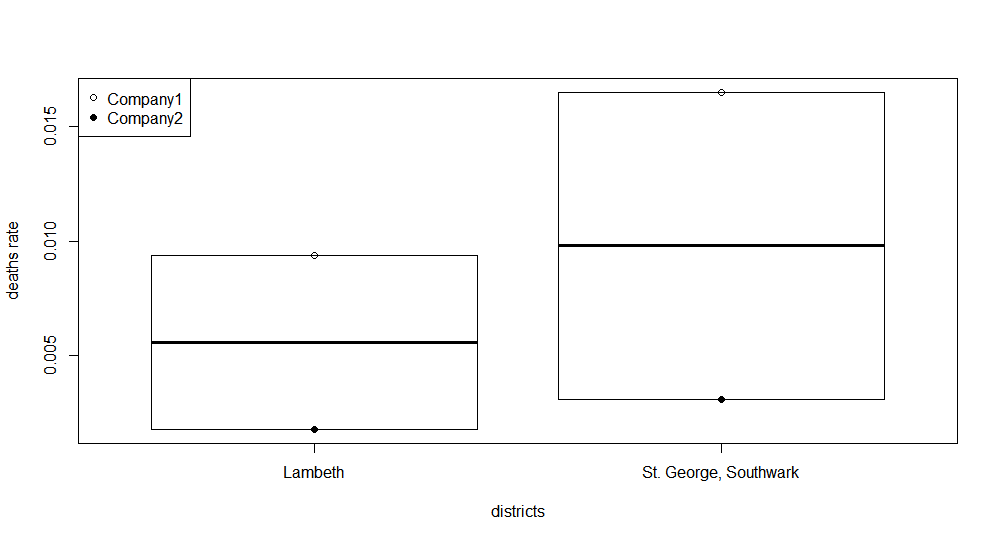
round(exp(coef(fit)),3)



summary(fit)



1. 調整company1為company2後，**St. Savior, Southwark**的deaths rate多0.021倍 (p-value<2e-16) < α(0.05)
2. 在相同地區下，調整company1為**company2**後，deaths rate多0.187倍 (p-value<2e-16) < α(0.05)
3. 調整company1為company2後， **Bermondsey**較St. Savior, Southwark的deaths rate多0.657倍 (p-value=3.99e-13) < α(0.05)
4. 調整company1為company2後， **Campberwell**較St. Savior, Southwark的deaths rate多0.705倍 (p-value=3.52e-07) < α(0.05)
5. 調整company1為company2後，**Lambeth**較St. Savior, Southwark的deaths rate多0.437倍 (p-value<2e-16) < α(0.05)
6. 調整company1為company2後，**Newington**較St. Savior, Southwark的deaths rate多0.629倍 (p-value=3.04e-13) < α(0.05)
7. 調整company1為company2後，**Rotherhithe**較St. Savior, Southwark的deaths rate多0.645倍 (p-value=1.51e-07) < α(0.05)
8. 調整company1為company2後，**St. George, Southwark**較St. Savior, Southwark的deaths rate多0.770倍 (p-value=4.91e-05) < α(0.05)
9. 調整company1為company2後，**St. Olave, Southwark**較St. Savior, Southwark的deaths rate多0.693倍 (p-value=1.25e-06) < α(0.05) < α(0.05)
10. 調整company1為company2後，**Wandsworth**較St. Savior, Southwark的deaths rate多0.670倍 (p-value=1.34e-07) < α(0.05)
11. 請畫出不同自來水公司之間地區St. George, Southwark及Lambeth分別對霍亂死亡率的mean response。



Code:

cholera\_noNA$deaths\_py<-cholera\_noNA$deaths/cholera\_noNA$py

cholera\_noNA\_sub<-subset(cholera\_noNA,cholera\_noNA$district=="St. George, Southwark"|cholera\_noNA$district=="Lambeth")

plot(cholera\_noNA$district,cholera\_noNA$deaths\_py,type="n")

points(cholera\_noNA\_sub$district,cholera\_noNA\_sub$deaths\_py)

attach(cholera\_noNA\_sub)

meanresponse<-cbind(

predict(fit,data.frame(district="St. George, Southwark",company="1",py=1),type="response"),

predict(fit,data.frame(district="St. George, Southwark",company="2",py=1),type="response"),

predict(fit,data.frame(district="Lambeth",company="1",py=1),type="response"),

predict(fit,data.frame(district="Lambeth",company="2",py=1),type="response")

)

meanresponse\_num<-meanresponse[1,]

district\_mean<-c("St. George, Southwark","St. George, Southwark","Lambeth","Lambeth")

company\_mean<-c("1","2","1","2")

meanresponseframe<-data.frame(district\_mean,company\_mean,meanresponse\_num)

plot(meanresponseframe$district\_mean,meanresponseframe$meanresponse\_num,xlab="districts",ylab="deaths rate")

points(meanresponseframe$district\_mean,meanresponseframe$meanresponse\_num,pch=c(1,16,1,16))

legend("topleft",legend=c("Company1","Company2"),pch=c(1,16))

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

加分題：

請利用y與x分別寫出(1)~(3)的模型，並說明y在什麼資料型態下才適用。(連續、類別、計數)

1. Simple linear regression 連續

Y=β0+βTX+𝜀, 𝜀 ~N( 0 , σ2 )

E[Y|X] = β0+βTX

1. logistic regression 類別

Y|X ~ Bernoulli(px) px =P(Y=1|X)=E[Y|X]

logit(px) = = β0+βTX

px =

(3) poisson regression 計數

Y|X ~ Poisson(m·λx) with m·λx = E[Y|X]

ln(λx)=β0+βTX

λx =