Statistics for Biology\_HW2

2017100057/ 이영노

2023-04-20

# 4.1

# (a)  
library(dplyr)

##   
## 다음의 패키지를 부착합니다: 'dplyr'

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

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

res = factor(c(-1,-1,-1,-1,1,1,1,1))%>%relevel(1);  
before = factor(c(-1,-1,1,1,-1,-1,1,1))%>%relevel(1);  
arrest = factor(c(-1,1,-1,1,-1,1,-1,1))%>%relevel(1);  
freq = c(75, 17, 109, 42, 359, 53, 175, 33);  
dat1=data.frame(res,before,arrest,freq);  
#tab1=xtabs(freq~.,dat1);  
  
fit = glm(arrest~res+before,weights=freq, family=binomial, data=dat1)  
summary(fit)

##   
## Call:  
## glm(formula = arrest ~ res + before, family = binomial, data = dat1,   
## weights = freq)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7 8   
## -5.853 7.349 -8.216 10.542 -9.762 14.866 -8.003 10.864   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.3603 0.1962 -6.933 4.13e-12 \*\*\*  
## res1 -0.5920 0.1968 -3.008 0.00263 \*\*   
## before1 0.3469 0.1906 1.820 0.06875 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 781.41 on 7 degrees of freedom  
## Residual deviance: 765.32 on 5 degrees of freedom  
## AIC: 771.32  
##   
## Number of Fisher Scoring iterations: 5

* Goodness-of-fit test for logistic regression model H0: B1=B2=0  
  Ha: not H0

LRT test statistic : ~  
where qchisq(0.975,2) is .  
Since LRT statistic > qchisq(0.975,2), reject H0 which means logistic model is not fit to data.

Even though logistic regression model is not a statistically significant model,  
we will go through the model as question said.

1. 로지스틱 회귀모형의 적합도 검정 실시 결과, 귀무가설을 기각할 수 없음에 따라 H0: B1=B2=0이다. 즉 로지스틱 회귀모형이 데이터와 fit하지 않는다.

절도피의자가 구속되는 로그오즈는 거주여부의 profile에 따라 -0.5920만큼, 체포경력 여부의 profile에 따라 0.3469 만큼 변화한다.

1. 거주민인 경우 그렇지 않은 경우보다 구속될 경우의 로그오즈가 2\*(-0.5920) 차이난다. 즉 거주민인 경우 구속될 오즈는 거주민이 아닐경우 구속될 오즈의 30.6%가 된다.

체포경력이 있는 경우 그렇지 않을 경우보다 구속될 경우의 로그오즈가 2\*(0.3469) 차이난다. 즉 체포경력이 있는 경우 구속될 오즈는 체포경력이 없는 경우 구속될 오즈의 200.13%가 된다.

# 4.3

# 명목형 반응변수 : generalized logit  
library(VGAM)

## Warning: 패키지 'VGAM'는 R 버전 4.2.3에서 작성되었습니다

## 필요한 패키지를 로딩중입니다: stats4

## 필요한 패키지를 로딩중입니다: splines

data=read.table("C:/Users/아로미/Downloads/Alligators2.dat",header=TRUE)  
class(data) # data.frame

## [1] "data.frame"

data$lake = as.factor(rep(c("Hancock","Oklawaha","Trafford","George"),rep(2,4)))  
fit=vglm(formula=cbind(y1,y2,y3,y4,y5) ~ lake + size, family=multinomial(refLevel='y1'),data=data)  
#lake가 범주가 4갠데 숫자로 되어있으므로, 가변수 취급하여 만들어줌. profile은 1,0으로 자동세팅됨.  
summary(fit)

##   
## Call:  
## vglm(formula = cbind(y1, y2, y3, y4, y5) ~ lake + size, family = multinomial(refLevel = "y1"),   
## data = data)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 -1.549019 0.424922 -3.645 0.000267 \*\*\*  
## (Intercept):2 -3.314533 1.053084 NA NA   
## (Intercept):3 -2.093077 0.662236 -3.161 0.001574 \*\*   
## (Intercept):4 -1.904272 0.525825 -3.621 0.000293 \*\*\*  
## lakeHancock:1 -1.658359 0.612877 -2.706 0.006813 \*\*   
## lakeHancock:2 1.242777 1.185432 1.048 0.294466   
## lakeHancock:3 0.695118 0.781263 0.890 0.373608   
## lakeHancock:4 0.826196 0.557540 1.482 0.138378   
## lakeOklawaha:1 0.937219 0.471906 1.986 0.047030 \*   
## lakeOklawaha:2 2.458872 1.118128 2.199 0.027871 \*   
## lakeOklawaha:3 -0.653208 1.202098 -0.543 0.586861   
## lakeOklawaha:4 0.005653 0.776513 0.007 0.994191   
## lakeTrafford:1 1.121985 0.490513 2.287 0.022174 \*   
## lakeTrafford:2 2.935253 1.116409 2.629 0.008559 \*\*   
## lakeTrafford:3 1.087767 0.841669 1.292 0.196221   
## lakeTrafford:4 1.516369 0.621435 2.440 0.014683 \*   
## size:1 1.458205 0.395945 3.683 0.000231 \*\*\*  
## size:2 -0.351263 0.580033 -0.606 0.544786   
## size:3 -0.630660 0.642480 -0.982 0.326296   
## size:4 0.331550 0.448247 0.740 0.459506   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1]),   
## log(mu[,4]/mu[,1]), log(mu[,5]/mu[,1])  
##   
## Residual deviance: 17.0798 on 12 degrees of freedom  
##   
## Log-likelihood: -47.5138 on 12 degrees of freedom  
##   
## Number of Fisher scoring iterations: 5   
##   
## Warning: Hauck-Donner effect detected in the following estimate(s):  
## '(Intercept):2'  
##   
##   
## Reference group is level 1 of the response

1. log(pi\_연체류/pi\_어류) = -1.549 -1.658*I(Hancock) + 0.937*I(Oklawaha) + 1.122*I(Trafford) + 1.458*I(Size)

log(pi\_파충류/pi\_어류) = -3.315 +1.243*I(Hancock) + 2.459*I(Oklawaha) + 2.935*I(Trafford) -0.351*I(Size)

log(pi\_조류/pi\_어류) = -2.093 +0.695*I(Hancock) -0.653*I(Oklawaha) + 1.088*I(Trafford) -0.631*I(Size)

log(pi\_기타/pi\_어류) = -1.904 +0.826*I(Hancock) + 0.006*I(Oklawaha) + 1.516*I(Trafford) + 0.332*I(Size)

어류 대비 ( A )를 먹이로서 선택할 추정로그오즈는 ( A )에 해당하는 식에, 해당하는 호수를 1로 나머지를 0을 대입, Size가 작으면 1 크면 0을 대입하여 나온 값이다.

예를들어 어류 대비 파충류를 먹이로서 선택할 추정로그오즈는 lake가 Hancock이고 Size가 2.3보다 작을때, -3.315+1.243-0.351 = -2.423이고, 추정오즈는 0.089 이다.

즉, Hancock호수에 사는 Size가 2.3보다 작은 악어는, 어류대비 파충류를 먹이로서 선택할 오즈가 8.9%로 어류 대비 파충류의 먹이 선택이 매우 적음을 알 수 있다.

(참고로, deviance는 17.07983이고, qchisq(0.975, 12)는 23.33666이다. deviance가 신뢰수준 95%수준에서 유의하지 않으므로, 귀무가설을 기각하지 못한다. H0: 모델이 데이터와 적합하다 이다. 즉, 일반화 로짓 모형은 해당 데이터를 설명하기에 신뢰수준 95% 에서 적합한 모델이다.)

library(dplyr)  
library(purrr)  
fct = function(lake, Size){  
 -1.549 -1.658\*as.numeric(lake=="Hancock") + 0.937\*as.numeric(lake=="Oklawaha") + 1.122\*as.numeric(lake=="Trafford") + 1.458\*as.numeric(Size>=2.3)}  
  
larger2.3 = c(fct("Hancock",2.5),fct("Oklawaha",2.6),fct("Trafford",6.6),fct("George",9.8))  
smaller2.3 = c(fct("Hancock",0.7),fct("Oklawaha",1),fct("Trafford",2.2),fct("George",0.2))  
result=data.frame(larger2.3, smaller2.3)  
or\_result = result%>%map\_df(exp)%>%data.frame()  
rownames(or\_result)=c("Hancock","Oklawaha","Trafford","George")  
print(or\_result)

## larger2.3 smaller2.3  
## Hancock 0.1739478 0.04047786  
## Oklawaha 2.3303070 0.54226525  
## Trafford 2.8038683 0.65246355  
## George 0.9130177 0.21246033

해당 테이블은 lake와 size에 따라 어류대비 연체류를 먹이로서 선택할 오즈비를 구한 것이다. 해당 테이블로 여러 해석이 가능한데,

예를들어 Hancock호수에 사는 size가 2.3보다 큰 악어가 어류대비 연체류를 먹이로서 선택할 오즈비는 0.1739이다. 즉, 해당 조건에 만족하는 악어가 어류대비 연체류를 먹이로서 선택할 확률은 17.39% 로서 (어류대비)연체류를 먹이로서 선택학 확률이 현저히 떨어짐을 알 수 있다.

반면 Okalawaha호수에 사는 size가 2.3보다 큰 악어가 어류대비 연체류를 먹이로서 선택할 오즈비는 2.33이다. 즉, 해당조건에 만족하는 악어가 어류대비 연체류를 먹이로서 선택할 확률은 233% 로서 (어류대비)연체류를 먹이로서 선택할 확률이 매우 높음을 알 수 있다.

해당 표로 다양한 해석이 가능한데, 특정 조건이 같을때 설명변수 하나의 변화에 대한 오즈비도 구할수 있다.

예를들어, lake가 Hancock으로 고정되어 있을때, Size가 큰 악어가 Size가 작은 악어보다 (어류대비)연체류를 먹이로서 선택할 오즈비는 4.2973이다.(exp(result[1,1]-result[1,2])) 즉, lake가 Hancock으로 동일할때 Size가 큰 악어가 Size가 작은악어 대비 연체류를 먹이로서 선택할 확률은 429.73% 로서, Hancock 호수에 사는 악어들은 Size가 클때 (어류대비)연체류를 먹이로서 선택할 확률이 매우 높다는 것을 알 수 있다.

또 예를들어, Size가 큰 악어로 고정되어 있을때, lake Okalawaha에 사는 악어가 lake Trafford에 사는 악어보다 (어류대비)연체류를 먹이로서 선택할 오즈비는 0.8311이다.(exp(result[2,1]-result[3,1])) 즉, size가 크다는 조건이 동일할때 lake Oklawaha에 사는 악어가 lake Trafford에 사는 악어대비 연체류를 먹이로서 선택할 확률은 83.11%로서, size가 큰 악어들은 lake Trafford대비 lake Okalawaha 호수에 살때 연체류를 먹이선택할 확률이 작다는 것을 알 수 있다.

# 4.4

library(tidyr)  
library(rje)

## Warning: 패키지 'rje'는 R 버전 4.2.3에서 작성되었습니다

##   
## 다음의 패키지를 부착합니다: 'rje'

## The following object is masked from 'package:VGAM':  
##   
## logit

## The following object is masked from 'package:dplyr':  
##   
## last

age = rep(factor(c("young","mid","old")),rep(3,1))%>%relevel("young","mid","old");  
test = rep(factor(c("often","sometimes","never")),rep(3,3))%>%relevel("often","sometimes","never");  
freq = c(91,150,109,90,200,198,51,155,172)  
dat4 = data.frame(age,test,freq)  
#tab4 = xtabs(freq~age+test)  
  
newdat4=dat4%>%spread(test,freq) # 세로로 된 데이터를 가로로 길게 만들기  
# newdat4$age=factor(c(1,2,3)) # 이렇게 코딩하는 거랑 동일한 결과값을 낸다.  
# 회귀식을 계산할때는 1,2,3 으로 코딩되지만, 결과를 해석할때는 회귀식이 indicator function으로   
# 나타내어지기 때문에 indicator function 에 따라 (1,0)으로 해석한다.  
# SAS profile 같이 여기서도 그런 코드가 있는지 찾아보았으나 찾지 못함.  
  
fit = vglm(cbind(never, sometimes, often) ~ age, family=cumulative(parallel=TRUE),  
 data=newdat4)  
summary(fit)

##   
## Call:  
## vglm(formula = cbind(never, sometimes, often) ~ age, family = cumulative(parallel = TRUE),   
## data = newdat4)  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 -1.2829 0.1296 -9.898 < 2e-16 \*\*\*  
## (Intercept):2 0.4470 0.1244 3.594 0.000326 \*\*\*  
## agemid 0.4419 0.1474 2.998 0.002715 \*\*   
## ageold 0.7315 0.1494 4.896 9.79e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])  
##   
## Residual deviance: 0.7124 on 2 degrees of freedom  
##   
## Log-likelihood: -18.6653 on 2 degrees of freedom  
##   
## Number of Fisher scoring iterations: 3   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
##   
## Exponentiated coefficients:  
## agemid ageold   
## 1.555663 2.078135

log[P(Y <= never|age\_i)/P(Y > never|age\_i)] = -1.2829 + 0.4419*age\_mid + 0.7315*age\_old

log[P(Y <= sometimes|age\_i)/P(Y > sometimes|age\_i)] = 0.4470 + 0.4419*age\_mid + 0.7315*age\_old

나이가 i일때 자가진단을 전혀 안하는 경우의 누적로그오즈는 해당 식으로 표현된다. 예를들어, 나이가 45세 미만일때는 age\_mid = 0, age\_old = 0 으로서 해당 식에 대입한 값이 된다.

해당 식을 통해 많은 것을 추정할수 있다. 예를들어, 45세 미만인 사람이 가끔 유방암 자가진단을 하는 확률을 구하고 싶을땐 다음과 같이 구할 수 있다.

P(Y <= never|age\_young)) = 0.2171 (expit(-1.2829 + 0.4419\*0 + 0.7315\*0)) P(Y <= sometimes|age\_young)) = 0.6099 (expit(0.4470 + 0.4419\*0 + 0.7315\*0))

Which implies,  
P(Y = never|age\_young) = 0.2171  
P(Y = sometimes|age\_young) = 0.6099 - 0.2171 = 0.3928  
P(Y = often|age\_young) = 1 - 0.6099 = 0.3901

따라서 45세 미만인 사람이 가끔 유방암 자가진단을 하는 확률은 39.28% 이다.

# 4.7

library(MASS)

##   
## 다음의 패키지를 부착합니다: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

belt = rep(factor(c("yes","no")),rep(4,2))%>%relevel("yes","no");  
eject = rep(rep(factor(c("yes","no")),rep(2,2)),2)%>%relevel("yes","no");  
mort = rep(factor(c("yes","no")),rep(4,1))%>%relevel("yes","no");  
freq = c(659, 270, 532, 347, 432, 532, 269, 552);  
dat7 = data.frame(belt,eject,mort,freq)  
  
fit = glm(freq ~ belt + eject + mort + belt:eject + belt:mort + eject:mort  
 ,family=poisson, data=dat7)  
summary(fit)

##   
## Call:  
## glm(formula = freq ~ belt + eject + mort + belt:eject + belt:mort +   
## eject:mort, family = poisson, data = dat7)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7 8   
## -0.09012 0.14136 0.10056 -0.12406 0.11163 -0.10027 -0.14081 0.09872   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.49423 0.03736 173.809 < 2e-16 \*\*\*  
## beltno -0.43118 0.05560 -7.755 8.82e-15 \*\*\*  
## ejectno -0.22195 0.05304 -4.185 2.86e-05 \*\*\*  
## mortno -0.90442 0.06205 -14.576 < 2e-16 \*\*\*  
## beltno:ejectno -0.23781 0.07011 -3.392 0.000694 \*\*\*  
## beltno:mortno 1.12236 0.07012 16.006 < 2e-16 \*\*\*  
## ejectno:mortno 0.48812 0.07017 6.956 3.51e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 330.21494 on 7 degrees of freedom  
## Residual deviance: 0.10569 on 1 degrees of freedom  
## AIC: 77.28  
##   
## Number of Fisher Scoring iterations: 3

fit$deviance > qchisq(0.975,1) # FALSE

## [1] FALSE

# so we do not reject H0, which means H0 model(부분연관모형) is better

0.10569 가 3.841 보다 작으므로, 3중교호작용 효과는 0이다.

model\_signif=function(x){  
 if (x-fit$deviance > qchisq(0.975,1)){  
 print("reject the null, which means lambda not zero")}  
 else{  
 print("null, which means lambda is zero")}}  
  
fit2 = glm(freq ~ belt + eject + mort + belt:eject + belt:mort, family=poisson, data=dat7)  
fit3 = glm(freq ~ belt + eject + mort + belt:eject + eject:mort, family=poisson, data=dat7)  
fit4 = glm(freq ~ belt + eject + mort + belt:mort + eject:mort, family=poisson, data=dat7)  
  
model\_signif(fit2$deviance)

## [1] "reject the null, which means lambda not zero"

model\_signif(fit3$deviance)

## [1] "reject the null, which means lambda not zero"

model\_signif(fit4$deviance)

## [1] "reject the null, which means lambda not zero"

1. 조건부 독립 결과들이 모두 significant하므로, 람다xy, 람다xz, 람다yz는 모두 0이 아니다. 따라서 최적의 모델은 fit 에 해당하는 부분연관모형 이다.

모든 변수들의 효과가 유의하게 나왔으며, 특히 효과가 크게 나타난 부분을 눈여겨볼만 하다.

사망한 사람에 비해 사망하지 않은 사람들의 오즈비가 0.4047로서 (exp(-0.90442)), 사망한 사람에 비해 사망하지 않은 사람들이 적다는 것을 의미한다.

벨트를 하고 죽은 사람들에 비해 벨트를 하지 않고 죽지 않은 사람들의 오즈비가 3.0721로서 (exp(1,12236)), 벨트를 하고 죽은 사람들에 비해 벨트를 하지 않고 죽지 않은 사람들이 매우 많다는 것을 의미한다.

fit = glm(mort ~ belt + eject , weights=freq, family=binomial, data=dat7)  
summary(fit)

##   
## Call:  
## glm(formula = mort ~ belt + eject, family = binomial, data = dat7,   
## weights = freq)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7 8   
## -21.16 25.92 -23.21 25.31 -26.42 25.06 -24.41 21.05   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.90442 0.06205 -14.576 < 2e-16 \*\*\*  
## beltno 1.12236 0.07012 16.006 < 2e-16 \*\*\*  
## ejectno 0.48812 0.07017 6.956 3.5e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4970.8 on 7 degrees of freedom  
## Residual deviance: 4663.8 on 5 degrees of freedom  
## AIC: 4669.8  
##   
## Number of Fisher Scoring iterations: 4

# LRT > qchisq so reject the null which means 교호작용이 0이 아니다.  
#fit$terms  
  
fit2 = glm(mort ~ belt + eject + belt:eject, weights = freq, family=binomial, data=dat7)  
summary(fit2)

##   
## Call:  
## glm(formula = mort ~ belt + eject + belt:eject, family = binomial,   
## data = dat7, weights = freq)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7 8   
## -21.27 25.83 -23.11 25.40 -26.33 25.15 -24.50 20.93   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.89230 0.07226 -12.349 < 2e-16 \*\*\*  
## beltno 1.10052 0.09703 11.342 < 2e-16 \*\*\*  
## ejectno 0.46498 0.09991 4.654 3.26e-06 \*\*\*  
## beltno:ejectno 0.04564 0.14038 0.325 0.745   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4970.8 on 7 degrees of freedom  
## Residual deviance: 4663.7 on 4 degrees of freedom  
## AIC: 4671.7  
##   
## Number of Fisher Scoring iterations: 4

벨트를 할때에 비해 벨트를 하지 않았을때 사망에 대한 오즈비는 3.005729 (exp(1.10052))로서 벨트를 할때에 비해 벨트를 하지 않았을때 사망할 확률이 더 높다.

차에서 튕겨나갔을때에 비해 차에서 튕겨나가지 않았을때 사망에 대한 오즈비는 1.5920 (exp(1.591982))로서 차에서 튕겨나갔을때에 비해 차에서 튕겨나가지 않았을때 사망할 확률이 더 높다.

벨트를 하고 차에서 튕겨나갔을때에 비해 벨트를 하지 않고 차에서 튕겨나가지 않았을 때 사망에 대한 오즈비는 유의하지 않으므로 유의미하게 해석할 수 없다.

# 4.8

class = rep(factor(c("I","II","III","IV")),rep(4,4))  
rent = rep(rep(factor(c("rent","own")),rep(2,2)),2)  
computer = rep(factor(c("yes","no")),rep(4,1))  
freq = c(304,38,92,64,665,85,174,113,894,93,379,321,720,84,433,297)  
dat8 = data.frame(class,rent,computer,freq)  
  
fit = glm(freq ~ class+rent+computer+class:rent+class:computer+rent:computer,  
 family=poisson, data =dat8)  
summary(fit)

##   
## Call:  
## glm(formula = freq ~ class + rent + computer + class:rent + class:computer +   
## rent:computer, family = poisson, data = dat8)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7 8   
## -0.1359 0.3939 0.2499 -0.2935 -0.3087 0.8957 0.6155 -0.7349   
## 9 10 11 12 13 14 15 16   
## 0.4886 -1.4345 -0.7370 0.8235 -0.1547 0.4615 0.2005 -0.2402   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.19534 0.10741 39.058 < 2e-16 \*\*\*  
## classII 0.60039 0.13045 4.603 4.17e-06 \*\*\*  
## classIII 1.52978 0.11727 13.045 < 2e-16 \*\*\*  
## classIV 1.51230 0.11803 12.813 < 2e-16 \*\*\*  
## rentrent -0.62234 0.11923 -5.220 1.79e-07 \*\*\*  
## computeryes 0.30027 0.12464 2.409 0.015991 \*   
## classII:rentrent 0.17051 0.12790 1.333 0.182484   
## classIII:rentrent -0.42503 0.11686 -3.637 0.000276 \*\*\*  
## classIV:rentrent -0.70527 0.11764 -5.995 2.03e-09 \*\*\*  
## classII:computeryes 0.01603 0.14655 0.109 0.912919   
## classIII:computeryes -0.05025 0.13440 -0.374 0.708526   
## classIV:computeryes 0.05317 0.13588 0.391 0.695570   
## rentrent:computeryes 1.85153 0.07789 23.772 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 3271.2182 on 15 degrees of freedom  
## Residual deviance: 5.9914 on 3 degrees of freedom  
## AIC: 146.09  
##   
## Number of Fisher Scoring iterations: 4

fit$deviance > qchisq(0.975,3) # so do not rejct the null which means 3중교호작용 is 0

## [1] FALSE

deviance 가 qchisq(0.975,3) 보다 작으므로, 3중교호작용 효과는 0이다.

fit2 = glm(freq ~ class+rent+computer+class:rent+class:computer,  
 family=poisson, data =dat8)  
fit3 = glm(freq ~ class+rent+computer+class:rent+rent:computer,  
 family=poisson, data =dat8)  
fit4 = glm(freq ~ class+rent+computer+class:computer+rent:computer,  
 family=poisson, data =dat8)  
  
model\_signif=function(x){  
 if (x-fit$deviance > qchisq(0.975,3)){  
 print("reject the null, which means lambda not zero")}  
 else{  
 print("null, which means lambda is zero")}}  
model\_signif(fit2$deviance)

## [1] "reject the null, which means lambda not zero"

model\_signif(fit3$deviance) # lambda(class:computer) is zero

## [1] "null, which means lambda is zero"

model\_signif(fit4$deviance)

## [1] "reject the null, which means lambda not zero"

람다jk(YZ) 가 0 이다. 우리는 산하에 있는 두개 모델을 검정할 수 있다.

fit5 = glm(freq ~ class+rent+computer+class:rent, family=poisson, data=dat8)  
fit6 = glm(freq ~ class+rent+computer+rent:computer, family=poisson, data=dat8)  
  
model\_signif\_nonZ=function(x){  
 if (x-fit3$deviance > qchisq(0.975,1)){  
 print("reject the null, which means lambda not zero")}  
 else{  
 print("null, which means lambda is zero")}}  
model\_signif\_Z=function(x){  
 if (x-fit3$deviance > qchisq(0.975,3)){  
 print("reject the null, which means lambda not zero")}  
 else{  
 print("null, which means lambda is zero")}}  
model\_signif\_nonZ(fit5$deviance)

## [1] "reject the null, which means lambda not zero"

model\_signif\_Z(fit6$deviance)

## [1] "reject the null, which means lambda not zero"

둘다 LRT statistic > qchisq. So reject the null which means 람다가 0이 아니다. 따라서 최적의 모델은 fit3에 해당하는 (XY,YZ) 모형이다.

summary(fit3)

##   
## Call:  
## glm(formula = freq ~ class + rent + computer + class:rent + rent:computer,   
## family = poisson, data = dat8)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.20193 -0.38907 0.01214 0.45586 1.36760   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.19290 0.08446 49.641 < 2e-16 \*\*\*  
## classII 0.60963 0.09947 6.129 8.86e-10 \*\*\*  
## classIII 1.50122 0.08854 16.956 < 2e-16 \*\*\*  
## classIV 1.54319 0.08820 17.495 < 2e-16 \*\*\*  
## rentrent -0.62089 0.11421 -5.436 5.44e-08 \*\*\*  
## computeryes 0.30452 0.04675 6.514 7.32e-11 \*\*\*  
## classII:rentrent 0.17564 0.11896 1.476 0.14   
## classIII:rentrent -0.44137 0.10852 -4.067 4.76e-05 \*\*\*  
## classIV:rentrent -0.68840 0.10931 -6.298 3.02e-10 \*\*\*  
## rentrent:computeryes 1.84840 0.07685 24.052 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 3271.2182 on 15 degrees of freedom  
## Residual deviance: 7.3992 on 6 degrees of freedom  
## AIC: 141.5  
##   
## Number of Fisher Scoring iterations: 3

classII : 사회적 지위가 1인 사람들에 비해 사회적 지위가 2인 사람들의 오즈비는 1.839751로서, 사회적지위가 1인 사람들에 비해 사회적 지위가 2인 사람들일 확률이 1.8397배 더 크다..

rentrent : 거주주택이 자가인 사람들에 비해 임대인 사람들의 오즈비는 0.5374로서, 주택이 자가인 사람들에 비해 임대인 사람들일 확률이 0.5374배 더 크다.

classII:rentrent : 사회적 지위가 1이고 거주주택이 자가인 사람들에 비해 사회적지위가 2이고 거주주택이 임대인 사람들의 오즈비는 1.192로서 사회적지위가 1이고 자가주택인 사람들에 비해 사회적지위가 2이고 임대주택인 사람들일 확률이 1.192배 더 크다. 그러나 이 해석은 유의수준 10%에서 유의하지 않다.

rentrent:computeryes : 거주주택이 자가이고 컴퓨터를 보유하지 않은 사람들에 비해 거주주택이 임대이고 컴퓨터를 보유한 사람들의 오즈비는 6.3497로서, 자가주택에 컴퓨터 미보유인 사람들에 비해 임대주택에 컴퓨터를 보유한 사람들일 확률이 6.3497배 더 크다.

# 5.3

library(car)

## Warning: 패키지 'car'는 R 버전 4.2.3에서 작성되었습니다

## 필요한 패키지를 로딩중입니다: carData

## Warning: 패키지 'carData'는 R 버전 4.2.3에서 작성되었습니다

##   
## 다음의 패키지를 부착합니다: 'car'

## The following object is masked from 'package:rje':  
##   
## logit

## The following object is masked from 'package:purrr':  
##   
## some

## The following object is masked from 'package:VGAM':  
##   
## logit

## The following object is masked from 'package:dplyr':  
##   
## recode

library(lsmeans)

## Warning: 패키지 'lsmeans'는 R 버전 4.2.3에서 작성되었습니다

## 필요한 패키지를 로딩중입니다: emmeans

## Warning: 패키지 'emmeans'는 R 버전 4.2.3에서 작성되었습니다

## The 'lsmeans' package is now basically a front end for 'emmeans'.  
## Users are encouraged to switch the rest of the way.  
## See help('transition') for more information, including how to  
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.

trt = rep(factor(c("A","B","C")),rep(10,3))  
x = c(11,8,5,14,19,6,10,6,11,3,6,6,7,8,18,8,19,8,5,15,16,13,11,9,21,16,12,12,7,12)  
y = c(6,0,2,8,11,4,13,1,8,0,0,2,3,1,18,4,14,9,1,9,13,10,18,5,23,12,5,16,1,20)  
dat5.3 = data.frame(trt,x,y)  
  
attach(dat5.3)

## The following objects are masked \_by\_ .GlobalEnv:  
##   
## trt, x, y

model0 = lm(y~trt\*x, data=dat5.3)  
Anova(model0,type=3)

## Anova Table (Type III tests)  
##   
## Response: y  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 5.08 1 0.3064 0.58499   
## trt 8.50 2 0.2566 0.77574   
## x 113.35 1 6.8427 0.01515 \*  
## trt:x 19.64 2 0.5930 0.56058   
## Residuals 397.56 24   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

trt\*x의 p\_Value가 0.56이므로 귀무가설을 기각할 수 없다. 따라서 교호작용이 존재하지 않으므로 공분산분석을 실행할 수 있다.

model1 = lm(y~trt+x,data=dat5.3)  
summary(model1)

##   
## Call:  
## lm(formula = y ~ trt + x, data = dat5.3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.4115 -2.3891 -0.5711 1.7237 8.5885   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.8808 1.9862 -1.954 0.0616 .   
## trtB 0.1090 1.7951 0.061 0.9521   
## trtC 3.4461 1.8868 1.826 0.0793 .   
## x 0.9872 0.1645 6.001 2.45e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.006 on 26 degrees of freedom  
## Multiple R-squared: 0.6763, Adjusted R-squared: 0.6389   
## F-statistic: 18.1 on 3 and 26 DF, p-value: 1.501e-06

Anova(model1,type=3)

## Anova Table (Type III tests)  
##   
## Response: y  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 61.26 1 3.8177 0.06155 .   
## trt 68.55 2 2.1361 0.13838   
## x 577.90 1 36.0145 2.454e-06 \*\*\*  
## Residuals 417.20 26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

trt의 p\_Value가 0.138이므로 귀무가설을 기각할 수 없다. 따라서 나병 치료약의 효과는 0으로 동일하다.

lsmeans(model1,~trt)

## trt lsmean SE df lower.CL upper.CL  
## A 6.71 1.29 26 4.07 9.36  
## B 6.82 1.27 26 4.21 9.44  
## C 10.16 1.32 26 7.46 12.87  
##   
## Confidence level used: 0.95

따라서 치료법 A,B,C 각각의 보정된 평균은 6.71, 6.82, 10.16이다.

# 5.5

trt = rep(factor(c(10,15,20)),rep(12,3))  
x1 = c(23.5,24,23,24.5,22.5,24.5,25.5,25,23,23,22.5,23,28.5,26,28.5,26.5,27,28,26.5,29,27,27,27,27,  
 25.6,23.1,20.8,19.2,21.3,26.6,25.5,18.9,29.3,30.2,18.9,25.3)  
x2 = c(15.0,15.5,14,13.5,13.5,13.5,13,14,14.5,14,14,15,16,15.5,17,16,18.5,14.5,17.5,13.5,16.5,16.5,  
 17.5,18.5,12,11.6,18.2,11.9,20.2,13.1,16.2,11.9,10.5,11.3,10.9,12.2)  
y = c(5,4.5,5,4.5,5.5,3,5,5,5.5,4.5,6.5,5.5,7,9,8.5,8.5,6,8,7,7,7,11,9,8,6.6,9.2,5.6,4.2,4.6,4.5,4.5,  
 6.1,6.3,8.1,4.9,6.2)  
dat5.5 = data.frame(trt,x1,x2,y)  
  
model0 = lm(y~trt\*x1+trt\*x2,data=dat5.5)  
Anova(model0,type=3)

## Anova Table (Type III tests)  
##   
## Response: y  
## Sum Sq Df F value Pr(>F)  
## (Intercept) 2.043 1 1.3006 0.2641  
## trt 3.339 2 1.0626 0.3596  
## x1 1.983 1 1.2621 0.2711  
## x2 0.004 1 0.0026 0.9597  
## trt:x1 5.533 2 1.7610 0.1911  
## trt:x2 0.138 2 0.0438 0.9572  
## Residuals 42.415 27

교호작용의 p\_Value가 x1, x2각각 0.1911, 0.9572 이므로, 귀무가설을 기각하지 못한다. 따라서 x1,x2각각의 교호작용은 0.05유의수준 하에 존재하지 않는다. 따라서 교호작용이 존재하지 않으므로 공분산분석을 실행할 수 있다.

model1 = lm(y ~ trt + x1 + x2, data=dat5.5)  
summary(model1)

##   
## Call:  
## lm(formula = y ~ trt + x1 + x2, data = dat5.5)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.09500 -0.84583 -0.00029 0.52904 3.08109   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.72951 3.06393 1.870 0.071 .   
## trt15 3.18230 0.70519 4.513 8.63e-05 \*\*\*  
## trt20 0.82294 0.52116 1.579 0.124   
## x1 0.05453 0.09395 0.580 0.566   
## x2 -0.14596 0.11107 -1.314 0.198   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.259 on 31 degrees of freedom  
## Multiple R-squared: 0.5595, Adjusted R-squared: 0.5027   
## F-statistic: 9.845 on 4 and 31 DF, p-value: 2.925e-05

Anova(model1,type=3)

## Anova Table (Type III tests)  
##   
## Response: y  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 5.541 1 3.4969 0.0709551 .   
## trt 32.433 2 10.2339 0.0003866 \*\*\*  
## x1 0.534 1 0.3369 0.5658462   
## x2 2.736 1 1.7268 0.1984655   
## Residuals 49.122 31   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

공변량을 제어한 제3종 제곱합의 결과에 따르면 물의양과 비료의양을 제어했을때 온도에 따라 성장속도가 유의하게 다르다. 그러나 공변량들의 효과는 유의수준 0.05 하에 유의하지 않다. 따라서 귀무가설을 기각 할수 없으므로, 각각의 공변량 x1, x2가 성장속도y에 미치는 영향이 없다고 판단된다.

1. 위에서 trt10을 0으로 놓았을때 trt15와의 차이가 유의하고, trt20과의 차이는 유의하지 않다고 결과값이 나왔다. 이를 tukey의 HSD다중비교를 통해 검정해보고자 한다.

options(contrasts=c("contr.sum","contr.poly"))  
model1.lsm = lsmeans(model1, pairwise~trt,glhargs=list())  
print(model1.lsm)

## $lsmeans  
## trt lsmean SE df lower.CL upper.CL  
## 10 4.95 0.390 31 4.15 5.75  
## 15 8.13 0.499 31 7.12 9.15  
## 20 5.77 0.417 31 4.92 6.62  
##   
## Confidence level used: 0.95   
##   
## $contrasts  
## contrast estimate SE df t.ratio p.value  
## trt10 - trt15 -3.182 0.705 31 -4.513 0.0002  
## trt10 - trt20 -0.823 0.521 31 -1.579 0.2696  
## trt15 - trt20 2.359 0.750 31 3.147 0.0099  
##   
## P value adjustment: tukey method for comparing a family of 3 estimates

tukey의 HSD검정을 시행해본 결과 trt10-trt15, trt15-trt20이 유의수준 0.05 하에서 유의한 차이가 있다.

# 5.7

trt = rep(factor(c("traditional","level","mix")),rep(20,3))  
x1 = c(99,95,99,99,102,101,102,103,102,103,102,104,106,106,107,108,109,111,110,113,95,97,100,100,  
 102,105,102,102,105,105,106,109,107,108,111,108,111,113,114,114,94,95,99,99,99,102,99,101,  
 104,101,104,107,107,106,109,107,109,111,111,112)  
x2 = c(12,11,13,13,12,12,12,13,13,14,14,13,15,13,13,12,13,13,13,13,12,12,13,13,13,12,13,14,12,13,  
 14,13,14,14,14,13,13,14,13,14,12,12,13,11,14,12,13,13,13,14,15,13,13,14,14,14,14,14,14,13)  
y = c(58,55,59,55,60,57,57,61,58,59,62,59,60,62,60,62,64,63,59,63,58,55,60,55,60,58,57,62,58,61,63,  
 59,61,64,61,63,65,64,62,60,59,56,62,57,61,60,58,63,60,63,65,61,62,64,61,65,63,66,62,64)  
dat5.7 = data.frame(trt,x1,x2,y)  
  
model0 = lm(y~trt\*x1 + trt\*x2, data=dat5.7)  
Anova(model0,type=3)

## Anova Table (Type III tests)  
##   
## Response: y  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 35.753 1 10.8406 0.001808 \*\*   
## trt 1.451 2 0.2200 0.803289   
## x1 84.997 1 25.7722 5.476e-06 \*\*\*  
## x2 39.624 1 12.0146 0.001080 \*\*   
## trt:x1 10.144 2 1.5379 0.224620   
## trt:x2 11.443 2 1.7348 0.186679   
## Residuals 168.199 51   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. 공변량 x1, x2 각각에 대하여 교호작용이 없으므로 공분산 분석을 시행할 수 있다.

model1 = lm(y~trt + x1 + x2, data=dat5.7)  
summary(model1)

##   
## Call:  
## lm(formula = y ~ trt + x1 + x2, data = dat5.7)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5603 -1.2775 0.0793 1.5272 3.3876   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.72031 5.12259 3.459 0.00105 \*\*   
## trt1 -0.61530 0.33812 -1.820 0.07423 .   
## trt2 1.10644 0.34259 3.230 0.00209 \*\*   
## x1 0.27747 0.05333 5.203 2.99e-06 \*\*\*  
## x2 1.05449 0.31766 3.320 0.00160 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.826 on 55 degrees of freedom  
## Multiple R-squared: 0.6004, Adjusted R-squared: 0.5713   
## F-statistic: 20.66 on 4 and 55 DF, p-value: 1.947e-10

공분산분석 결과를 보면 공변량 각각에 대하여 p\_Value가 2.99e-06, 0.00160으로 0.05수준 하에서 유의하다. 따라서 귀무가설(공변량의 효과가 0이다.)을 기각할 수 있으므로 공변량 각각의 효과가 유의수준 0.05 하에서 0이 아니다.

trt\_mix와 trt\_traditional는 trt\_level의 추정치가 0일때, trt\_level과의 차이를 나타낸다. trt\_mix의 효과는 유의수준 0.05 하에서 유의하지만, trt\_traditional은 유의하지 않다. 따라서 trt\_mix는 귀무가설(trt\_mix = 0)을 기각할 수 있으므로 trt\_mix의 효과가 0이 아니다. 그러나 trt\_traditional은 귀무가설(trt\_traditional = 0)을 기각할 수 없으므로, trt\_traditional의 효과가 0이다.

1. y\_mix = 17.10501 + 1.72174 + 0.27747*x1 + 1.05449*x2

y\_level = 17.10501 + 0.27747*x1 + 1.05449*x2

y\_traditional = 17.10501 + 0.12417 + 0.27747*x1 + 1.05449*x2