범주형자료분석방법론 기말고사

이름: 김연주

학과: 통계학과

학번: 2021250461

Q1

		Response to Chemotherapy			
Therapy	Gender	Progressive Disease	No Change	Partial Remission	Complete Remission
Alternating	Male	41	44	20	20
Alternating	Female	12	7	3	1
Sequential	Male	28	45	29	26
Sequential	Female	4	12	5	2

Obtain the fit of a baseline-category logit model with 'No Change' as the baseline category, and then answer the following questions.

(a) Report the fitted model(s) for the data.

아래와 같은 data1을 생성한 후 baseline category logit model인 fit1을 만들었다.

```
> data1 <- data.frame(
+   Therapy = rep(c("A", "S"), each = 8),
+   Gender = rep(c("Male", "Female"), each = 4, times = 2),
+   Response = rep(c(1, 2, 3, 4), times = 4), # 1=pd, 2=nc, 3=pr, 4=cr
+   Count = c(41, 44, 20, 20, 12, 7, 3, 1, 28, 45, 29, 26, 4, 12, 5, 2)
+ )
> data1
   Therapy Gender Response Count
1        A   Male        1    41
2        A   Male        2    44
3        A   Male        3    20
4        A   Male        4    20
5        A Female        1    12
6        A Female        2     7
7        A Female        3     3
8        A Female        4     1
9        S   Male        1    28
10       S   Male        2    45
11       S   Male        3    29
12       S   Male        4    26
13       S Female        1     4
14       S Female        2    12
15       S Female        3     5
16       S Female        4     2
```

이때, 두번째 범주인 'no change'를 baseline으로 두기 위해 반응변수를 범주형 변수로 설정하고 기준 범주를 nc (no change)로 설정하였다. Fit1의 summary 결과는 아래와 같다.

```
> # Response를 factor로 변환하고 기준 범주 설정
> data1$Response <- factor(data1$Response, levels = c(1, 2, 3, 4)) # 범주 설정
> data1$Response <- relevel(data1$Response, ref = "2") # '2' (No Change) 기준 설정
> # VGAM 패키지 로드
> library(VGAM)
> # Baseline-Category Logit 모델 적합
> fit1 <- vglm(Response ~ Therapy + Gender,
+              family = multinomial(refLevel = "2"),
+              weights = Count,
+              data = data1)
> # 모델 요약 출력
> summary(fit1)

Call:
vglm(formula = Response ~ Therapy + Gender, family = multinomial(refLevel = "2"),
    data = data1, weights = Count)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1   0.1126     0.3677    0.31   0.7595
(Intercept):2  -1.0261     0.4646   -2.21   0.0272 *
(Intercept):3  -1.9479     0.6528   -2.98   0.0028 **
TherapyS:1     -0.6165     0.2955   -2.09   0.0369 *
TherapyS:2      0.2814     0.3320    0.85   0.3965
TherapyS:3      0.1822     0.3492    0.52   0.6019
GenderMale:1   -0.0902     0.3786   -0.24   0.8117
GenderMale:2    0.2716     0.4580    0.59   0.5532
GenderMale:3    1.1881     0.6474    1.84   0.0665 .
---
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])

Residual deviance: 786 on 39 degrees of freedom

Log-likelihood: -393 on 39 degrees of freedom

Number of Fisher scoring iterations: 5

No Hauck-Donner effect found in any of the estimates


Reference group is level  1  of the response
>
```

위에서 기준 범주는 두번째 범주인 'nc'이다. 따라서 결과는 다른 반응변수들을 기준 범주와 비교한 로짓 비율에 대해 적합되었다. (Interpret):1은 첫번째 범주인 'pd'와 'nc'의 로짓 비율에 대한 절편값이다. 같은 방법으로 (interpret):2와 (interpret):3은 pr과 nc, cr과 nc의 로짓 비율에 대한 절편값이다.

(b) Interpret the therapy effect in the fitted model(s).

```
> # TherapyS 계수 추출
> therapy_effect <- coefficients[grep("TherapyS", names(coefficients))]
> therapy_effect
TherapyS:1 TherapyS:2 TherapyS:3
   -0.6165     0.2814     0.1822
>
```

1) Progressive disease: TherapyS:1의 계수가 -0.6165이며 통계적으로 유의하다. 이는 pd 대비 nc의 로그 오즈가 A보다 S에서 0.6165 낮음을 의미한다. Exp(-0.6165)=0.54이므로 A에 비해 S의 pd확률이 46% 낮아진다.
2) Partial Remission: TherapyS:2의 계수가 0.2814이며 통계적으로 유의하지 않다. 이는 pr 대비 nc의 로그 오즈가 A보다 S에서 0.2814 높음을 의미한다. Exp(0.2814)=1.32이므로 A에 비해 S의 pd확률이 32% 증가한다.
3) Complete remission: TherapyS:3의 계수가 0.1822이며 통계적으로 유의하지 않다. 이는 cr 대비 nc의 로그 오즈가 A보다 S에서 0.1822 높음을 의미한다. Exp(0.1822)=1.2이므로 A에 비해 S의 pd확률이 20% 높아진다.

(c) Test the significance of model parameters.

therapyS:1을 제외하고 회귀계수의 z검정통계량이 모두 유의하지 않다. 이때 귀무가설은 해당 회귀계수가 통계적으로 유의하지 않다는 것이고, p-value가 0.05보다 크면 귀무가설을 기각하지 못한다.

```
> # 잔차 편차와 자유도 확인
> residual_deviance <- deviance(fit1) # Residual Deviance
> df_residual <- df.residual(fit1)   # 자유도
> # 잔차 편차를 기준으로 p-value 계산
> p_value <- pchisq(residual_deviance, df_residual, lower.tail = FALSE)
> # 결과 출력
> cat("Residual Deviance:", residual_deviance, "\n")
Residual Deviance: 786
> cat("Degrees of Freedom:", df_residual, "\n")
Degrees of Freedom: 39
> cat("p-value:", p_value, "\n")
p-value: 7.748e-140
```

(d) Estimate the probability that a male patient with the sequential therapy has a complete remission.

Fit1의 회귀계수를 통해 각 반응변수의 범주별 logit값을 계산할 수 있다. Pd의 logit=0.1126−0.6165−0.0902=−0.5941, pr logit=−1.0261+0.2814+0.2716=−0.4731, cr logit=−1.9479+0.1822+1.1881=−0.5776, nc의 logit은 기준 범주이기에 0이 된다.

$$P(Y=j)=\frac{\exp(\text{logit}_j)}{\sum_{k=1}^{4}\exp(\text{logit}_k)}$$

위 식을 통해 각 범주의 확률을 계산할 수 있다. 이때 cr (complete remission)의 확률은 0.5611/2.7367=0.2050이다. 즉, Therapy=S, gender=Male의 조건 하에서 cr의 확률은 20.5%이다.

Obtain the fit of a cumulative logit model, and then answer the following questions.

(e) Report the fitted model(s) for the data.

```
> # Response를 순서형 변수로 변환
> data1$Response = ordered(data1$Response, levels = c(1, 2, 3, 4))  # 1=pd, 2=nc, 3=pr,
4=cr
> # Cumulative Logit Model 적합
> fit1_cum = vglm(Response ~ Therapy + Gender, family = cumulative(parallel = TRUE),  w
eights = Count, data = data1)
> summary(fit1_cum)

Call:
vglm(formula = Response ~ Therapy + Gender, family = cumulative(parallel = TRUE),
    data = data1, weights = Count)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1   -0.196      0.295   -0.67   0.5060
(Intercept):2    1.371      0.306    4.48  7.4e-06 ***
(Intercept):3    2.422      0.328    7.39  1.4e-13 ***
TherapyS        -0.581      0.212   -2.74   0.0061 **
GenderMale      -0.541      0.295   -1.83   0.0667 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]), logitlink(P[Y<=3])

Residual deviance: 789.1 on 43 degrees of freedom

Log-likelihood: -394.5 on 43 degrees of freedom

Number of Fisher scoring iterations: 4

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
  TherapyS GenderMale
    0.5595     0.5819
>
```

반응변수를 순서형 변수로 변환한 후 cumulative logit model인 fir2를 생성하였다. 이때 'pd'는 가장 낮은 범주, 'cr'은 가장 높은 범주이다. 누적 로짓은 아래와 같다.

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_1 \cdot \text{Therapy} + \beta_2 \cdot \text{Gender}.$$

이때 절편 1,2,3은 Therapy, Gender가 기준값 (therapy=A, gender=Female)일 때 각 누적 반응 수준의 로그 오즈값을 의미한다. 또한, TherapyS의 계수인 -0.581을 통해 Therapy가 S일 때 P(Y>j)의 로그 오즈가 A일 때보다 0.581 낮으며, S일 경우 P(Y<=j)의 오즈가 A의 exp(-0.581)=0.5595배라는 것을 알 수 있다. GenderM의 계수를 통해서도 위와 비슷한 해석을 할 수 있다. 남성인 경우에 여성보다 P(Y<j)의 로그오즈가 0.541 낮고, P(Y<=j)의 오즈가 여성의 exp(-0.541)=0.5819배이다.

(f) Interpret the therapy effect in the fitted model(s).

TherapyS의 계수는 -0.581이다. 이는 A일 때에 비해 S일 때 누적 확률 P(Y<=j)에 대한 로그 오즈가 0.581 감소함을 의미한다. 오즈비는 exp(-0.581)=0.5595로, 이는 S일 때 A에 비해 누적 확률의 오즈가 44% 감소됨을 보여준다.

(g) Test the significance of model parameters.

GenderMale의 회귀계수를 제외하고 모두 z검정 결과 통계적으로 유의하였다. 또한, GenderMale의 회귀계수의 p-value도 0.05에 가깝다.

(h) Estimate the probability that a male patient with the sequential therapy has a complete remission.

Fit2의 회귀계수와 아래의 식을 통해 각 반응변수의 범주별 cumulative logit값을 계산할 수 있다.

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_{\text{TherapyS}} \cdot \text{Therapy} + \beta_{\text{GenderMale}} \cdot \text{Gender}.$$

계산 결과는 다음과 같다.

1. $P(Y \leq 1)$ (Progressive Disease 이하):

$$\text{logit}[P(Y \leq 1)] = \alpha_1 + (-0.581) + (-0.541) = -0.196 - 0.581 - 0.541 = -1.31$$

$$P(Y \leq 1) = \frac{\exp(-1.318)}{1 + \exp(-1.318)} \approx 0.211.$$

2. $P(Y \leq 2)$ (No Change 이하):

$$\text{logit}[P(Y \leq 2)] = \alpha_2 + (-0.581) + (-0.541) = 1.371 - 0.581 - 0.541 = 0.249.$$

$$P(Y \leq 2) = \frac{\exp(0.249)}{1 + \exp(0.249)} \approx 0.562.$$

3. $P(Y \leq 3)$ (Partial Remission 이하):

$$\text{logit}[P(Y \leq 3)] = \alpha_3 + (-0.581) + (-0.541) = 2.422 - 0.581 - 0.541 = 1.300.$$

$$P(Y \leq 3) = \frac{\exp(1.300)}{1 + \exp(1.300)} \approx 0.786.$$

이때 각 반응변수의 범주별 확률은 $P(Y = j) = P(Y \leq j) - P(Y \leq j - 1)$ 의 식으로 계산 가능하기에 complete remission의 확률은 $P(Y=4)=1-P(Y\leq 3)=1-0.786=0.214$이다.

Q2

Quality of Management (Q)	Supervisor's Job Satisfaction (S)	Worker's Job Satisfaction (W)		
		Low	High	Total
Good	Low	59	109	168
	High	78	205	283
Bad	Low	103	87	190
	High	32	42	74

Fit loglinear models to the data, and then answer the following questions.

(a) Use the likelihood ratio statistic to test the null hypothesis of no three-factor interaction.

Data2 생성 후 full_model과 reduced_model (no three-factor model)을 만들었다.

```
> # 데이터 입력
> data2 = array(c(59, 109, 78, 205, 103, 87, 32, 42),
+               dim = c(2, 2, 2),
+               dimnames = list(
+                 Worker = c("Low", "High"),
+                 Supervisor = c("Low", "High"),
+                 Quality = c("Good", "Bad")
+               ))
> data2
, , Quality = Good

      Supervisor
Worker Low High
  Low   59   78
  High 109  205

, , Quality = Bad

      Supervisor
Worker Low High
  Low  103   32
  High  87   42
```

```
> summary(full_model)
Formula:
~Quality * Supervisor * Worker
attr(,"variables")
list(Quality, Supervisor, Worker)
attr(,"factors")
           Quality Supervisor Worker Quality:Supervisor Quality:Worker
Quality          1          0      0                  1              1
Supervisor       0          1      0                  1              0
Worker           0          0      1                  0              1
           Supervisor:Worker Quality:Supervisor:Worker
Quality                    0                         1
Supervisor                 1                         1
Worker                     1                         1
attr(,"term.labels")
[1] "Quality"                   "Supervisor"
[3] "Worker"                    "Quality:Supervisor"
[5] "Quality:Worker"            "Supervisor:Worker"
[7] "Quality:Supervisor:Worker"
attr(,"order")
[1] 1 1 1 2 2 2 3
attr(,"intercept")
[1] 1
attr(,"response")
[1] 0
attr(,".Environment")
<environment: R_GlobalEnv>

Statistics:
               X^2 df P(> X^2)
Likelihood Ratio   0  0        1
Pearson            0  0        1
```

```
> # Reduced Model (No three-factor interaction)
> reduced_model <- loglm(~ (Quality + Supervisor + Worker)^2, data = data2)
> # 모델 결과 출력
> summary(reduced_model)
Formula:
~(Quality + Supervisor + Worker)^2
attr(,"variables")
list(Quality, Supervisor, Worker)
attr(,"factors")
           Quality Supervisor Worker Quality:Supervisor Quality:Worker
Quality          1          0      0                  1              1
Supervisor       0          1      0                  1              0
Worker           0          0      1                  0              1
           Supervisor:Worker
Quality                    0
Supervisor                 1
Worker                     1
attr(,"term.labels")
[1] "Quality"            "Supervisor"         "Worker"
[4] "Quality:Supervisor" "Quality:Worker"     "Supervisor:Worker"
attr(,"order")
[1] 1 1 1 2 2 2
attr(,"intercept")
[1] 1
attr(,"response")
[1] 0
attr(,".Environment")
<environment: R_GlobalEnv>

Statistics:
                     X^2 df P(> X^2)
Likelihood Ratio 0.06494  1   0.7989
Pearson          0.06488  1   0.7989
```

```
> # LRT
> anova(reduced_model, full_model, test = "Chisq")
LR tests for hierarchical log-linear models

Model 1:
 ~(Quality + Supervisor + Worker)^2
Model 2:
 ~Quality * Supervisor * Worker

          Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1    0.06494  1
Model 2    0.00000  0    0.06494         1         0.7988
Saturated  0.00000  0    0.00000         0         1.0000
>
```

귀무가설은 reduced_model이 데이터를 잘 적합한다는 것으로 세 요인 간의 상호작용이 없는 것으로 해석 가능하다. LRT 실행 결과 p-value가 0.7988로 귀무가설이 채택된다. 즉, 세 요인 간 상호작용이 없다는 가정이 데이터를 잘 설명한다. 세 요인 간 상호작용을 포함하지 않아도 데이터 적합도에 큰 차이가 없기 때문이다.

(b) Use the likelihood ratio statistics for goodness of fits of the (QS, QW), (QW, SW), and (SW, QS) models. Explain why the (QS, QW, SW) model is the best.

```
> # (QS, QW) 모델 적합
> fit2.2 = loglm(~ Quality + Supervisor + Worker + Quality:Supervisor + Quality:Worker,
data2)
> # (QW, SW) 모델 적합
> fit2.3 = loglm(~ Quality + Supervisor + Worker + Quality:Worker + Supervisor:Worker,
data2)
> # (SW, QS) 모델 적합
> fit2.4 = loglm(~ Quality + Supervisor + Worker + Supervisor:Worker + Quality:Supervis
or, data2)
> # (QS, QW, SW) 모델 적합
> fit2.5 = loglm(~ Quality + Supervisor + Worker + Quality:Supervisor + Quality:Worker
+ Supervisor:Worker, data2)

> summary(fit2.2)
Formula:
~Quality + Supervisor + Worker + Quality:Supervisor + Quality:Worker
attr(,"variables")
list(Quality, Supervisor, Worker)
attr(,"factors")
           Quality Supervisor Worker Quality:Supervisor Quality:Worker
Quality          1          0      0                  1              1
Supervisor       0          1      0                  1              0
Worker           0          0      1                  0              1
attr(,"term.labels")
[1] "Quality"            "Supervisor"         "Worker"
[4] "Quality:Supervisor" "Quality:Worker"
attr(,"order")
[1] 1 1 1 2 2
attr(,"intercept")
[1] 1
attr(,"response")
[1] 0
attr(,".Environment")
<environment: R_GlobalEnv>

Statistics:
                  X^2 df P(> X^2)
Likelihood Ratio 5.387  2  0.06764
Pearson          5.410  2  0.06686
```

(QS, QW) model인 fit2.2의 LRT 결과 귀무가설 (해당 모델이 데이터를 잘 설명한다)이 채택되어 적합도가 높음을 알 수 있다.

```
> summary(fit2.3)
Formula:
~Quality + Supervisor + Worker + Quality:Worker + Supervisor:Worker
attr(,"variables")
list(Quality, Supervisor, Worker)
attr(,"factors")
           Quality Supervisor Worker Quality:Worker Supervisor:Worker
Quality          1          0      0              1                 0
Supervisor       0          1      0              0                 1
Worker           0          0      1              1                 1
attr(,"term.labels")
[1] "Quality"           "Supervisor"        "Worker"            "Quality:Worker"
[5] "Supervisor:Worker"
attr(,"order")
[1] 1 1 1 2 2
attr(,"intercept")
[1] 1
attr(,"response")
[1] 0
attr(,".Environment")
<environment: R_GlobalEnv>

Statistics:
                  X^2 df  P(> X^2)
Likelihood Ratio 19.71  2 5.245e-05
Pearson          19.88  2 4.811e-05
```

(QW, SW) model인 fit2.3의 LRT 결과 귀무가설 (해당 모델이 데이터를 잘 설명한다)이 기각되어 적합도가 낮음을 알 수 있다.

```
> summary(fit2.4)
Formula:
~Quality + Supervisor + Worker + Supervisor:Worker + Quality:Supervisor
attr(,"variables")
list(Quality, Supervisor, Worker)
attr(,"factors")
           Quality Supervisor Worker Supervisor:Worker Quality:Supervisor
Quality          1          0      0                 0                  1
Supervisor       0          1      0                 1                  1
Worker           0          0      1                 1                  0
attr(,"term.labels")
[1] "Quality"            "Supervisor"         "Worker"
[4] "Supervisor:Worker"  "Quality:Supervisor"
attr(,"order")
[1] 1 1 1 2 2
attr(,"intercept")
[1] 1
attr(,"response")
[1] 0
attr(,".Environment")
<environment: R_GlobalEnv>

Statistics:
                  X^2 df  P(> X^2)
Likelihood Ratio 71.90  2 2.220e-16
Pearson          70.88  2 4.441e-16
```

(SW, QS) model인 fit2.4의 LRT 결과 귀무가설 (해당 모델이 데이터를 잘 설명한다)이 기각되어 적합도가 낮음을 알 수 있다.

```
> anova(fit2.2, fit2.5)
LR tests for hierarchical log-linear models

Model 1:
 ~Quality + Supervisor + Worker + Quality:Supervisor + Quality:Worker
Model 2:
 ~Quality + Supervisor + Worker + Quality:Supervisor + Quality:Worker + Supervisor:Work
er

          Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1    5.38714  2
Model 2    0.06494  1    5.32220         1        0.02106
Saturated  0.00000  0    0.06494         1        0.79885
> anova(fit2.3, fit2.5)
LR tests for hierarchical log-linear models

Model 1:
 ~Quality + Supervisor + Worker + Quality:Worker + Supervisor:Worker
Model 2:
 ~Quality + Supervisor + Worker + Quality:Supervisor + Quality:Worker + Supervisor:Work
er

          Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   19.71112  2
Model 2    0.06494  1   19.64618         1        0.00001
Saturated  0.00000  0    0.06494         1        0.79885
```

```
> anova(fit2.4, fit2.5)
LR tests for hierarchical log-linear models

Model 1:
 ~Quality + Supervisor + Worker + Supervisor:Worker + Quality:Supervisor
Model 2:
 ~Quality + Supervisor + Worker + Quality:Supervisor + Quality:Worker + Supervisor:Work
er

          Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   71.90156  2
Model 2    0.06494  1   71.83662         1         0.0000
Saturated  0.00000  0    0.06494         1         0.7988
>
```

Fit2.5와 fit2.2, 2.3, 2.4를 비교한 결과 세 모델에서 모두 reduced model이 적합하다는 귀무가설이 기각되어 (QS, QW, SW) 모델이 데이터를 더욱 잘 적합한다는 결과가 도출되었다.

```
> summary(fit2.5)
Formula:
~Quality + Supervisor + Worker + Quality:Supervisor + Quality:Worker +
    Supervisor:Worker
attr(,"variables")
list(Quality, Supervisor, Worker)
attr(,"factors")
           Quality Supervisor Worker Quality:Supervisor Quality:Worker
Quality          1          0      0                  1              1
Supervisor       0          1      0                  1              0
Worker           0          0      1                  0              1
           Supervisor:Worker
Quality                    0
Supervisor                 1
Worker                     1
attr(,"term.labels")
[1] "Quality"            "Supervisor"         "Worker"
[4] "Quality:Supervisor" "Quality:Worker"     "Supervisor:Worker"
attr(,"order")
[1] 1 1 1 2 2 2
attr(,"intercept")
[1] 1
attr(,"response")
[1] 0
attr(,".Environment")
<environment: R_GlobalEnv>

Statistics:
                     X^2 df P(> X^2)
Likelihood Ratio 0.06494  1   0.7989
Pearson          0.06488  1   0.7989
```

Fit2.5의 LRT 결과에서도 해당 모델이 데이터를 잘 적합한다는 것을 알 수 있다.

(c) At a fixed level of management quality, compute the odds ratio at low and high levels of supervisor satisfaction using the fitted (QS, QW, SW) model. Interpret it.

```
> expected <- fitted(fit2.5)
Re-fitting to get fitted values
> expected
, , Quality = Good

      Supervisor
Worker    Low   High
  Low   59.73  77.26
  High 108.27 205.74

, , Quality = Bad

      Supervisor
Worker    Low  High
  Low  102.27 32.74
  High  87.73 41.26
```

1) management quality=good: supervisor= low일 오즈는 59.73/108.27=0.552, high일 오즈는 0.375 이므로 오즈비는 0.681이다. 이는 supervisor의 만족도가 low에서 high로 증가하면 worker 가 low일 오즈가 0.68배가 됨을 의미한다.

2) management quality=bad: supervisor=high일 오즈는 102.27/87.74=1.165, low일 오즈는 0.793 이므로 오즈비는 0.681이다. 이는 supervisor의 만족도가 low에서 high로 증가하면 worker 가 low일 오즈가 0.68배가 됨을 의미한다. 이처럼 두 오즈비의 값은 동일하며 supervisor의 만족도가 high일 때 worker의 만족도가 low일 오즈는 supervisor의 만족도가 low일 때의 0.681배임을 알 수 있다.

```
> # Odds Ratio 계산 함수
> compute_odds_ratio <- function(expected, quality) {
+   # Supervisor = Low
+   odds_low <- expected["Low", "Low", quality] / expected["High", "Low", quality]
+
+   # Supervisor = High
+   odds_high <- expected["Low", "High", quality] / expected["High", "High", quality]
+
+   # Odds Ratio
+   odds_ratio <- odds_high / odds_low
+   return(list(odds_low = odds_low, odds_high = odds_high, odds_ratio = odds_ratio))
+ }
> # Quality = Good
> result_good <- compute_odds_ratio(expected, "Good")
> cat("Quality = Good:\n")
Quality = Good:
> cat("  Odds (Supervisor = Low):", result_good$odds_low, "\n")
  Odds (Supervisor = Low): 0.5517
> cat("  Odds (Supervisor = High):", result_good$odds_high, "\n")
  Odds (Supervisor = High): 0.3755
> cat("  Odds Ratio:", result_good$odds_ratio, "\n\n")
  Odds Ratio: 0.6807

> # Quality = Bad
> result_bad <- compute_odds_ratio(expected, "Bad")
> cat("Quality = Bad:\n")
Quality = Bad:
> cat("  Odds (Supervisor = Low):", result_bad$odds_low, "\n")
  Odds (Supervisor = Low): 1.166
> cat("  Odds (Supervisor = High):", result_bad$odds_high, "\n")
  Odds (Supervisor = High): 0.7934
> cat("  Odds Ratio:", result_bad$odds_ratio, "\n")
  Odds Ratio: 0.6807
```

(d) Show a loglinear-logit model connection using the (QS, QW, SW) model.

Loglinear model은 셀 빈도를 기반으로 변수 간 상호작용을 모델링한다. 즉, 이는 셀 빈도의 로그값을 주효과와 상호작용 효과의 합으로 표현한다. 반면, logit model은 특정 이항 반응변수의 로그 오즈를 설명한다. 이때, liglinear 모델에서 특정 반응변수 Y에 대한 로그 오즈는 다른 변수들의 주 효과 및 상호작용 효과로 나타난다. 아래와 같은 로그 오즈 계산값이 logit model의 식과 동일한 것이다.

$$\log\left(\frac{\mu_{i1k}}{\mu_{i2k}}\right) = (\lambda_Y^1 - \lambda_Y^2) + (\lambda_{XY}^{i1} - \lambda_{XY}^{i2}) + (\lambda_{YZ}^{1k} - \lambda_{YZ}^{2k})$$

$$\log\left(\frac{P(Y=1|X,Z)}{P(Y=0|X,Z)}\right) = \alpha + \beta_X + \beta_Z.$$

즉, loglinear model에서 상호작용 계수는 logit 모델의 상호작용 계수와 동일한 정보를 제공한다. 이러한 이유로 아래의 코드에서 두 모델의 log_odds의 값이 거의 동일한 것을 확인할 수 있다.

```
> # 로그선형 모델 적합
> fit_loglinear <- loglm(~ Quality + Supervisor + Worker +
+                            Quality:Supervisor + Quality:Worker + Supervisor:Worker, d
ata2)
>
> # 예상값 추출
> expected <- fitted(fit_loglinear)
Re-fitting to get fitted values
>
> # 로그 오즈 계산 (Good vs Bad)
> log_odds_loglinear <- log(expected[,, "Good"] / expected[,, "Bad"])
> cat("Log Odds (Loglinear Model):\n")
Log Odds (Loglinear Model):
> print(log_odds_loglinear)
      Supervisor
Worker     Low    High
  Low  -0.5946  0.1532
  High -0.9793 -0.2316
```

```
> # 로짓 모델 적합
> fit_logit <- glm(cbind(Freq_Good, Freq_Bad) ~ Supervisor * Worker,
+                  family = binomial(link = "logit"),
+                  data = data_logit)
>
> # 로짓 모델 계수 기반 로그 오즈 계산
> logit_coeff <- coef(fit_logit)
> log_odds_logit <- matrix(NA, nrow = 2, ncol = 2, dimnames = list(c("Low", "High"), c
("Low", "High")))
>
> for (w in c("Low", "High")) {
+     for (s in c("Low", "High")) {
+         intercept <- logit_coeff["(Intercept)"]
+         sup_effect <- ifelse(s == "High", logit_coeff["SupervisorHigh"], 0)
+         work_effect <- ifelse(w == "High", logit_coeff["WorkerHigh"], 0)
+         interaction_effect <- ifelse(s == "High" & w == "High", logit_coeff["Supervis
orHigh:WorkerHigh"], 0)
+
+         log_odds_logit[w, s] <- intercept + sup_effect + work_effect + interaction_ef
fect
+     }
+ }
>
> cat("Log Odds (Logit Model):\n")
Log Odds (Logit Model):
> print(log_odds_logit)
         Low    High
Low  -0.6138  0.1688
High -0.9663 -0.2719
```

```
> # 로그 오즈 차이 확인
> cat("Difference between Loglinear and Logit Model Log Odds:\n")
Difference between Loglinear and Logit Model Log Odds:
> print(log_odds_loglinear - log_odds_logit)
       Supervisor
Worker       Low     High
  Low    0.01922 -0.01564
  High -0.01304  0.04037
>
```

Q3

	MI Cases		
MI Controls	Diabetes	No Diabetes	Total
No Diabetes	37	82	119
Diabetes	9	16	25
Total	46	98	144

(a) Test whether the marginal proportions are dependent or not, and interpret the result.

```
> # McNemar's Test
> mcnemar.test(data3)

        McNemar's Chi-squared test with continuity correction

data:  data3
McNemar's chi-squared = 7.5, df = 1, p-value = 0.006
```

귀무가설은 두 변수의 marginal proportion이 서로 독립이라는 것이다. 즉, 데이터에서 두 변수 간의 통계적으로 유의한 차이가 없다는 의미이다. 맥니마 검정 결과 p-value가 매우 낮아 귀무가설이 기각된다. 즉, MI와 당뇨 발병은 통계적으로 종속 관계에 있다.

(b) Obtain the 95% confidence interval for the difference of marginal proportions, and interpret the result.

```
> n=9+16+37+82
> d=(16-37)/n
> se=sqrt(16+37)/n
> d+1.96*se
[1] -0.04674
> d-1.96*se
[1] -0.2449
```

한계비율 차이 d=-0.1458로, 이는 MI Cases에서 당뇨병 상태로 전환된 비율이 대조군에서보다 약 14.58% 낮음을 의미한다.

신뢰구간은 [-0.04674, -0.2449]로, 0을 포함하지 않기에 한계 비율의 차이는 통계적으로 유의하다. 즉, 두 변수는 통계적으로 종속 관계를 나타낸다.

Q4

Responses			Gender			
			Female		Male	
First	Second	Third	Junior	Senior	Junior	Senior
Not Obese	Not Obese	Not Obese	11	29	33	11
Not Obese	Not Obese	Obese	7	8	19	23
Not Obese	Obese	Not Obese	8	7	45	67
Not Obese	Obese	Obese	3	9	43	12
Obese	Not Obese	Not Obese	11	23	13	6
Obese	Not Obese	Obese	4	2	39	16
Obese	Obese	Not Obese	11	7	29	17
Obese	Obese	Obese	16	14	19	19

(a) Fit an appropriate model to the data, and report the fitted model.

```
> head(data4)
      First    Second     Third Gender AgeGroup Count
1 Not Obese Not Obese Not Obese Female   Junior    11
2     Obese Not Obese Not Obese Female   Junior     7
3 Not Obese     Obese Not Obese Female   Junior     8
4     Obese     Obese Not Obese Female   Junior     3
5 Not Obese Not Obese     Obese Female   Junior    11
6     Obese Not Obese     Obese Female   Junior     4
> # GEE 모델 적합
> library(geepack)
> data4$First = ifelse(data4$First == "Not Obese", 0, 1)
> gee_model <- geeglm(First ~ Second + Third + Gender + AgeGroup,
+                     family = binomial(link = "logit"),
+                     data = data4,
+                     id = interaction(Gender, AgeGroup),
+                     weights = Count)  # Count를 가중치로 설정
```

```
> # GEE 모델 적합
> library(geepack)
> data4$First = ifelse(data4$First == "Not Obese", 0, 1)
> gee_model <- geeglm(First ~ Second + Third + Gender + AgeGroup,
+                     family = binomial(link = "logit"),
+                     data = data4,
+                     id = interaction(Gender, AgeGroup),
+                     weights = Count)  # Count를 가중치로 설정
> summary(gee_model)

Call:
geeglm(formula = First ~ Second + Third + Gender + AgeGroup,
    family = binomial(link = "logit"), data = data4, weights = Count,
    id = interaction(Gender, AgeGroup))

 Coefficients:
             Estimate Std.err   Wald Pr(>|W|)
(Intercept)   -0.6022  0.1459  17.05  3.6e-05 ***
SecondObese   -0.2387  0.5436   0.19  0.66057
ThirdObese     0.6612  0.1791  13.63  0.00022 ***
GenderMale    -0.3128  0.0291 115.66  < 2e-16 ***
AgeGroupSenior  0.4684  0.1231  14.48  0.00014 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)        1  0.0351
Number of clusters:   4  Maximum cluster size: 8
```

(b) Test the significance of model parameters, and interpret model parameters.

Model의 summary 결과 회귀계수의 Wald 통계량이 유의하지 않은 변수는 Second (Obese)가 유일했다. 즉, 해당 변수를 제외한 다른 변수들의 회귀계수는 모두 통계적으로 유의하였다.

Intercept: -0.6022는 기준 그룹 (여성, junior, second와 third 모두 not obese)의 로그 오즈값이다. 이를 확률로 변환하면 기준 그룹의 obese 확률은 exp(-0.6022)/(1+exp(-0.6022))=0.35이다.

Third (Obese): 0.6612는 Third 시점에서 Obese였던 상태가 First 시점에서 Obese일 확률을 유의미하게 증가시킴을 의미한다. 이를 확률로 변환하면 obese 확률의 증가 비율은 exp(0.6612)=1.94이다. 즉, Third 시점에 Obese였을 때 First 시점의 Obese 확률이 약 94% 증가한다.

Gender (Male): -0.3128는 남성이 여성에 비해 Obese일 확률이 유의하게 낮음을 보여준다. 이를 확률로 변환하면 남성의 비만 오즈비는 exp(-0.3128)=0.73이다. 즉, 남성은 여성보다 비만일 확률이 27% 낮다.

AgeGroup (Senior): 0.4684는 Senior 그룹이 Junior 그룹에 비해 비만일 확률이 유의하게 증가함을 보여준다. 이를 확률로 변환하면 Senior의 비만 오즈비는 exp(0.4684)=1.6이다. 즉, Senior은 Junior그룹보다 비만일 확률이 60% 높다.