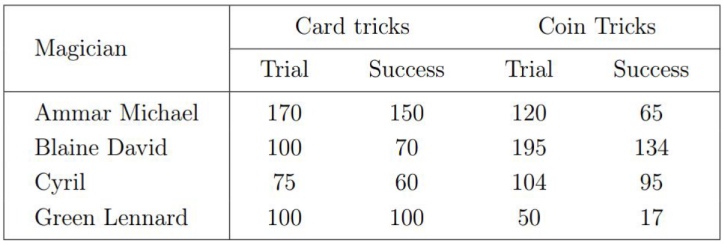
Final Project for STA5003 Categorical Data Analysis

12032925 刘艺璇

> This report is based on two problems: the magician data and the belief data, analysis and code are both provided for illustartion.

1. Here is the data for 4 magicians. The table shows the type of tricks, the number of tricks performed and number of successful tricks.



Magician	Card tricks		Coin Tricks	
	Trial	Success	Trial	Success
Ammar Michael	170	150	120	65
Blaine David	100	70	195	134
Cyril	75	60	104	95
Green Lennard	100	100	50	17

(a) Fit a saturated logistic model for the successful rate of a trick. Try to use both effect coding (1 and 0) model and reference coding (1 and -1) model.

- Treat the successful rate of a trick as π, in this problem, dummy variables are introduced in order to distinguish between the **magician individuals** (Ammar Michael, Blaine David, Cyril and Green Lennard), and **trick type** (card tricks or coin tricks).

  The saturated logistic regression model is constructed by :

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1-\pi(x)} = \alpha + \mathbf{X_1}\mathrm{B_1} + \mathbf{X_2}\mathrm{B_2} + \mathbf{X_1^T}\mathbf{X_2}\mathrm{B_3}$$

where $\pi(x) = P(Y=1|X=x)$, $\mathbf{x_1}$ is the indicator matrix for magician individuals, $\mathrm{B_1}$ is the coefficient vector for each dummy variable of magicians; $\mathbf{x_2}$ is the indicator matrix for trick type, $\mathrm{B_2}$ is the coefficient vector for each dummy variable of trick type.

- First use **reference coding**:
  - For each $\mathbf{X_i}$, $i = 1, 2$, reference coding is defined to be:

$$\mathrm{X_{ij}} = \begin{cases} 1, & \text{if level j,} \\ 0, & \text{otherwise.} \end{cases}$$

  Where $j = 1, \cdots, k-1$. Set magician Ammar Michael as $\mathbf{X_{11}} = (0,0,0)$, Blaine David as $\mathbf{X_{12}} = (1,0,0)$ and so on. $\mathbf{X_2}$ is a number instead of a vector in this case. Set the card trick as $x_2 = 1$, and the coin tricks as $x_2 = 0$.

```
dmyr<-dummyVars(~Magician,data=magic_df,fullRank=T)
dummy_r<-data.frame(predict(dmyr,newdata=magic_df))
dummy_r=rbind(dummy_r,dummy_r)
Tricks.Card=c(rep(1,4),rep(0,4))
magic_rc['Trial']=c(magic_df$CardTrial,magic_df$CoinTrial)
magic_rc['Success']=c(magic_df$CardSuccess,magic_df$CoinSuccess)
magic_rc['Fail']=magic_rc['Trial']-magic_rc['Success']
Tricks=magic_rc[c('Success','Fail')]
```

By using the sequel above, the table after reference coding is as followed:

	Magician.Blaine.David	Magician.Cyril	Magician.Green.lennard	Tricks.Card	Trial	Success	Fail
Ammar Michael	0	0	0	1	170	150	20
Blaine David	1	0	0	1	100	70	30
Cyril	0	1	0	1	75	60	15
Green lennard	0	0	1	1	100	100	0
Ammar Michael1	0	0	0	0	120	65	55
Blaine David1	1	0	0	0	195	134	61
Cyril1	0	1	0	0	104	95	9
Green lennard1	0	0	1	0	50	17	33

- Consider a **saturated logistic regression model** with **intersections** between magician individuals and trick type. By using the sequel below:

```
m1=glm(as.matrix(Tricks)~as.matrix(dummy_r)*Tricks.Card,family=binomial)
summary(m1)
```

```
Call:
glm(formula = as.matrix(Tricks) ~ as.matrix(dummy_r) * Tricks.Card,
    family = binomial)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
                                                       Estimate Std. Error z
value Pr(>|z|)
(Intercept)                                              0.1671     0.1832
0.912  0.36187
as.matrix(dummy_r)Magician.Blaine.David                  0.6199     0.2396
2.587  0.00968 **
as.matrix(dummy_r)Magician.Cyril                         2.1896     0.3940
5.558 2.73e-08 ***
as.matrix(dummy_r)Magician.Green.lennard                -0.8303     0.3503
-2.371  0.01776 *
Tricks.Card                                              1.8478     0.3004
6.152 7.67e-10 ***
as.matrix(dummy_r)Magician.Blaine.David:Tricks.Card     -1.7875     0.4021
-4.445 8.78e-06 ***
as.matrix(dummy_r)Magician.Cyril:Tricks.Card            -2.8182     0.5433
-5.187 2.14e-07 ***
as.matrix(dummy_r)Magician.Green.lennard:Tricks.Card    26.1266 51688.8861
0.001  0.99960
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.6211e+02  on 7  degrees of freedom
Residual deviance: 2.7542e-10  on 0  degrees of freedom
AIC: 48.983

Number of Fisher Scoring iterations: 22
```

The **saturated logistic regression model** (M_1) could be further written as

$$\mathrm{logit}[\pi(x)] = 0.1671 + \mathbf{X_1}\begin{bmatrix}0.6199\\2.1896\\-0.8303\end{bmatrix} + 1.8478\mathbf{x_2} + \mathbf{X_1^T x_2}\begin{bmatrix}-1.7875\\-2.8182\\26.1266\end{bmatrix}$$

- Then use **effect coding**:
  - Another way to impose constraints sets $\sum_j \mathrm{B}_{ij} = 0$. For each $\mathbf{x_i}$, $i = 1, 2$, effect coding is defined to be:

$$\mathbf{X_{ij}} = \begin{cases} 1, & \text{if level j,}\\ -1, & \text{if level k,}\\ 0, & \text{otherwise.}\end{cases}$$

Where $j = 1, \cdots, k-1$. Set magician Ammar Michael as $\mathbf{X_{11}} = (1,0,0)$, Blaine David as $\mathbf{X_{12}} = (0,1,0)$, Cyril as $\mathbf{X_{13}} = \mathbf{(0,0,1)}$ and $\mathbf{X_{14}} = (-1,-1,-1)$. $\mathbf{x_2}$ is a number instead of a vector in this case. Set the card trick as $x_2 = 1$, and the coin tricks as $x_2 = -1$.

In this case, use the sum coding technique by the ***contr.sum(4)*** sequel in R language, as the sequel below.

```
  dummy_r1<-contr.sum(4)
  dimnames(dummy_r1)=list(c("Ammar Michael","Blaine David","Cyril","Green lennard"))
  Magicians<-rbind(dummy_r1,dummy_r1)
  Tricks.Card1=c(rep(1,4),rep(-1,4))
  dummy2<-cbind(Magicians,Tricks)
  dummy2['Tricks.Card']=Tricks.Card1
```

The table of data used in the following regression is shown:

	1	2	3	Success	Fail	Tricks.Card
Ammar.Michael	1	0	0	150	20	1
Blaine.David	0	1	0	70	30	1
Cyril	0	0	1	60	15	1
Green.lennard	-1	-1	-1	100	0	1
Ammar.Michael.1	1	0	0	65	55	-1
Blaine.David.1	0	1	0	134	61	-1
Cyril.1	0	0	1	95	9	-1
Green.lennard.1	-1	-1	-1	17	33	-1

- Then, again Consider a **saturated logistic regression model** with **intersections** between magician individuals and trick type. By using the sequel below:

```
m2=glm(as.matrix(Tricks)~as.matrix(Magicians)*Tricks.Card1,family=binomial)
summary(m2)
```

```
Call:
glm(formula = as.matrix(Tricks) ~ as.matrix(Magicians) * Tricks.Card1,
    family = binomial)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                           4.276   6461.111   0.001    0.999
as.matrix(Magicians)1                -3.185   6461.111   0.000    1.000
as.matrix(Magicians)2                -3.459   6461.111  -0.001    1.000
as.matrix(Magicians)3                -2.404   6461.111   0.000    1.000
Tricks.Card1                          3.614   6461.111   0.001    1.000
as.matrix(Magicians)1:Tricks.Card1   -2.690   6461.111   0.000    1.000
as.matrix(Magicians)2:Tricks.Card1   -3.584   6461.111  -0.001    1.000
as.matrix(Magicians)3:Tricks.Card1   -4.099   6461.111  -0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.6211e+02  on 7  degrees of freedom

Residual deviance: 2.7541e-10  on 0  degrees of freedom
AIC: 48.983

Number of Fisher Scoring iterations: 22
```

The **saturated logistic regression model** (M_2) could be further written as:

$$\text{logit}[\pi(x)] = 4.276 + \mathbf{X_1}\begin{bmatrix}-3.185\\-3.459\\-2.404\end{bmatrix} + 3.614\mathbf{x_2} + \mathbf{X_1^T x_2}\begin{bmatrix}-2.690\\-3.584\\-4.099\end{bmatrix}$$

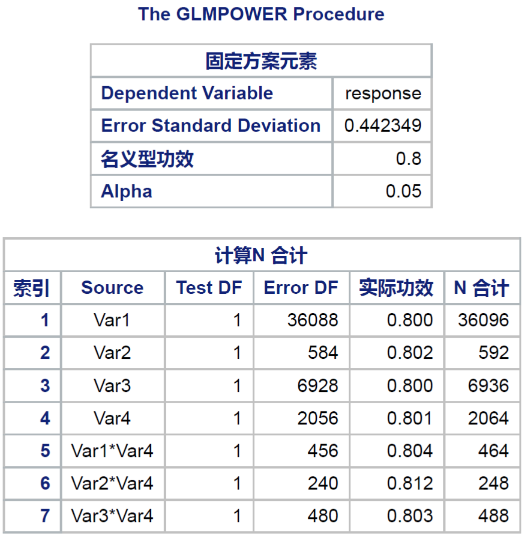
However, from the regression result it has very large standard errors, which is probably caused by the collinearity when introducing intersections. This indicates that the estimations are very unstable.

(b) Comment on the adequacy of the sample size.

- Intuitively speaking, when two codings are applied for a saturated model, M_1 has mostly significant coefficients while M_2 doesn't. This is possibly caused by the imbalance sample of magician Green lennard play cards and the lack of sample size, which leads to the failure when fitting M_2.
- Similar to the sample size and power consideration from the lecture notes, consider **power analysis** for determining the optimal sample size. Tried to achieve through **SAS Proc GLMPOWER**. Where the response rate is the expected ratio of the number of success over number of trials, `STDDEV =0.04423488` which corresponds to `sqrt(p(1-p))` where p is the average of the shown response rates. Suppose we want a power of 0.8, then estimation is done through following sequel:

```
data exemplar;
  input Var1 $ Var2 $ Var3 $ Var4 $ response;
  datalines;
    0 0 0 1 0.8824
    1 0 0 1 0.7000
    0 1 0 1 0.8000
    0 0 1 1  0.9999
    0 0 0 0 0.5417
    1 0 0 0  0.6872
    0 1 0 0 0.9135
    0 0 1 0 0.3400000
run;

proc glmpower data=exemplar;

  class Var1 Var2 Var3 Var4;
  model response = Var1 Var2 Var3 Var4 Var1*Var4 Var2*Var4 Var3*Var4;
  power
    power=0.8
    ntotal=.
    stddev=0.4423488;
run;
```

Then the output is as followed:



The GLMPOWER Procedure

固定方案元素	
Dependent Variable	response
Error Standard Deviation	0.442349
名义型功效	0.8
Alpha	0.05

计算N 合计

索引	Source	Test DF	Error DF	实际功效	N 合计
1	Var1	1	36088	0.800	36096
2	Var2	1	584	0.802	592
3	Var3	1	6928	0.800	6936
4	Var4	1	2056	0.801	2064
5	Var1*Var4	1	456	0.804	464
6	Var2*Var4	1	240	0.812	248
7	Var3*Var4	1	480	0.803	488

Where we need 36096 sample size as our sample size, since $Var1$ main effect is the hardest to estimate, with power equal to 0.80 and alpha equal to 0.05. This is way larger than 1605 we have.

(c) Base on the fitted model in (a), determine the estimated successful rates for all magician trick combinations.

The main idea of estimated the successful rates are through $\text{logit}[\hat{\pi}\left(x_0\right)] = \hat{\alpha} + \hat{\beta}x_0$, then

$$\hat{\pi}\left(x_0\right) = \frac{exp(\hat{\alpha} + \hat{\beta}x_0)}{1 + exp(\hat{\alpha} + \hat{\beta}x_0)}$$

- For model M_1, if use the ***predict*** function, the successful rate is shown for all magicians and trick combinations.

  ```
  predict(m1,type = "response")
  ```

  The estimated successful rate for all combinations is shown as followed.

	Ammar Michael	Blaine David	Cyril	Green Lennard
Card Trick	0.8823529	0.7000000	0.8000000	1.0000000
Coin Trick	0.5416667	0.6871795	0.9134615	0.3400000

which is exactly as the observed value, it makes sense for the saturated model.

- For model M_2, use the predict function again.

  ```
  predict(m2,type="response")
  ```

  The estimated rate for all combinations is shown as followed. It is exactly the same as previous estimation, this is also a saturated model.

	Ammar Michael	Blaine David	Cyril	Green Lennard
Card Trick	0.8823529	0.7000000	0.8000000	1.0000000
Coin Trick	0.5416667	0.6871795	0.9134615	0.3400000

(d) Find the odds ratio of Cyril's card trick compared to Green's coin trick as well as its 95% confidence interval.

In order to conduct odds ratio estimation and confidence interval calculation for Cyril's crad trick and Green's coin trick, first subtract the subtable from the sequel as followed:

```
rate=matrix(c(60,15,17,33),nrow=2,byrow=TRUE)
dimnames(rate)=list(c("Cyril","Green"),c("Sucecss","Fail"))
```

The table is shown:

	Success	Fail
Cyril	60	15
Green	17	33

The ratio of odds ratio is calculated by $\theta_1/\theta_2 = \frac{\pi_1(x)(1-\pi_2(x))}{\pi_2(x)(1-\pi_2(x))} = \frac{n_{11}n_{22}}{n_{12}n21} = 7.764706$, the 95% confidence interval is shown as followed, with other measures also included for comparison.

```
rate=matrix(c(60,15,17,33),nrow=2,byrow=TRUE)
dimnames(rate)=list(c("Cyril","Green"),c("Sucecss","Fail"))
Wald.ci<-function(Table, aff.response, alpha=.05){
  # Gives two-sided Wald CI's for odds ratio, difference in proportions and
relative risk.
  # Table is a 2x2 table of counts with rows giving the treatment populations
  # aff.response is a string like "c(1,1)" giving the cell of the beneficial
response and the
  # treatment category
  # alpha is significance level
  pow<-function(x, a=-1) x^a
  z.alpha<-qnorm(1-alpha/2)
```

```
  if(is.character(aff.response))
    where<-eval(parse(text=aff.response))
  else where<-aff.response

  Next<-as.numeric(where==1) + 1

  # OR
  odds.ratio<-
  Table[where[1],where[2]]*Table[Next[1],Next[2]]/(Table[where[1],Next[2]]*Table
[Next[1],where[
      2]])
  se.OR<-sqrt(sum(pow(Table)))
  ci.OR<-exp(log(odds.ratio) + c(-1,1)*z.alpha*se.OR)

  # difference of proportions
  p1<-Table[where[1],where[2]]/(n1<-Table[where[1],Next[2]] +
Table[where[1],where[2]])
  p2<-Table[Next[1],where[2]]/(n2<-
Table[Next[1],where[2]]+Table[Next[1],Next[2]])

  se.diff<-sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
  ci.diff<-(p1-p2) + c(-1,1)*z.alpha*se.diff

  # relative risk
  RR<-p1/p2
  se.RR<-sqrt((1-p1)/(p1*n1) + (1-p2)/(p2*n2))
  ci.RR<-exp(log(RR) + c(-1,1)*z.alpha*se.RR)

  list(OR=list(odds.ratio=odds.ratio, CI=ci.OR),
proportion.difference=list(diff=p1-p2,

 CI=ci.diff), relative.risk=list(relative.risk=RR,CI=ci.RR))
}
wald.ci(rate, "c(1, 1)")
```

The result is provided as followed, when the critical value $\alpha = 0.05$, the confidence interval for the ratio of odds ratio is:

```
$OR
$OR$odds.ratio
[1] 7.764706

$OR$CI
[1]  3.440611 17.523242


$proportion.difference
$proportion.difference$diff
[1] 0.46

$proportion.difference$CI
[1] 0.3005146 0.6194854


$relative.risk
$relative.risk$relative.risk
[1] 2.352941
```

```
$relative.risk$CI
[1] 1.573407 3.518689
```

(e) Fit a reduced logistic model with only main effects (i.e., no interaction). Obtain confidence intervals for coefficients. Comment on the goodness of fit.

- For both coding techniques, considering only the main effect model, then the logistic regression model could be considered as:

$$\operatorname{logit}[\pi(x)] = \log \frac{\pi(x)}{1-\pi(x)} = \alpha + \mathbf{X_1}\mathrm{B_1} + \mathbf{X_2}\mathrm{B_2}$$

  where $\pi(x) = P(Y=1|X=x)$, $\mathbf{x_1}$ is the indicator matrix for magician individuals, B_1 is the coefficient vector for each dummy variable of magicians; $\mathbf{x_2}$ is the indicator matrix for trick type, B_2 is the coefficient vector for each dummy variable of trick type.

  - For the reference coding technique, the model is denoted as M_3.

```
m3=glm(as.matrix(Tricks)~as.matrix(dummy_r)+Tricks.Card,family=binomial)
summary(m3)
```

```
Coefficients:
                                          Estimate Std. Error z value
Pr(>|z|)
(Intercept)                                0.45737    0.15932   2.871
0.00409 **
as.matrix(dummy_r)Magician.Blaine.David    0.02234    0.19340   0.116
0.90802
as.matrix(dummy_r)Magician.Cyril           1.03975    0.26494   3.924
8.69e-05 ***
as.matrix(dummy_r)Magician.Green.lennard   0.12518    0.24589   0.509
0.61069
Tricks.Card                                1.14559    0.17365   6.597
4.19e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model has three significant coefficients and two insignificant from the p-value.

  - For the effect coding technique, the model is denoted as M_4.

```
m4=glm(as.matrix(Tricks)~as.matrix(Magicians)+Tricks.Card1,family=binomi
al)
summary(m4)
```

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)             1.32698    0.09218  14.395  < 2e-16 ***
as.matrix(Magicians)1 -0.29682    0.13325  -2.228   0.0259 *  
as.matrix(Magicians)2 -0.27447    0.12956  -2.119   0.0341 *  
as.matrix(Magicians)3  0.74293    0.18216   4.078 4.53e-05 ***
Tricks.Card1            0.57279    0.08682   6.597 4.19e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model returns a even more significant set of coefficients, compared with model M_2 that has strong multicollinearity.

- Then consider the 95% **confidence interval** of coefficients

  For the Wald approach, the interval is calculated as $\hat{\beta} \pm Z_{\alpha/2}(SE)$

  - For the reference coding technique:

    ```
    confint(m3)
    ```

    Two insignificant coefficients of magician Cyril and magician green lennard, have confidence intervals containing 0, which again indicates that they are not significant.

    ```
                                                   2.5 %    97.5 %
    (Intercept)                                0.1474738 0.7730138
    as.matrix(dummy_r)Magician.Blaine.David   -0.3570027 0.4019253
    as.matrix(dummy_r)Magician.Cyril           0.5336420 1.5756826
    as.matrix(dummy_r)Magician.Green.lennard -0.3509750 0.6152741
    Tricks.Card                                0.8096412 1.4911474
    ```

  - For the effect coding technique:

    ```
    confint(m4)
    ```

    The 95% confidence interval of all coefficients do not contain 0, which coincides with the conclusion that they are all significant.

    ```
                               2.5 %      97.5 %
    (Intercept)            1.1504597  1.51229445
    as.matrix(Magicians)1 -0.5579601 -0.03492472
    as.matrix(Magicians)2 -0.5282290 -0.01981839
    as.matrix(Magicians)3  0.3984550  1.11520063
    Tricks.Card1           0.4048206  0.74557371
    ```

    In comparison with the confidence interval estimation from model M_2, the confidence interval is evidently smaller, indicating it provides a more robust result.

    The 95% confidence interval coefficients estimation of M_2 is:

    ```
                                          2.5 %     97.5 %
    (Intercept)                        3790.826  2744.3758
    as.matrix(Magicians)1             -2637.532 -1719.4700
    as.matrix(Magicians)2             -2716.878 -3267.6975
    as.matrix(Magicians)3             -2585.626  -717.8447
    Tricks.Card1                       4789.103  2794.7418
    as.matrix(Magicians)1:Tricks.Card1 -2760.823 -4142.2675
    as.matrix(Magicians)2:Tricks.Card1 -2698.377 -2903.1932
    as.matrix(Magicians)3:Tricks.Card1 -2710.390 -3128.7969
    ```

- Then, considered three types of **goodness of fit** for model compared with the **saturated model** : Pearson chi-squared (X^2), likelihood ratio (G^2) and Wald chi-squared. Consider the common GLM models with $Var(Y_i) = v(\mu_i)$ and $\phi(1) = 1$,

  Pearson residual is defined as:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\nu\left(\hat{\mu}_i\right)}}$$

Then, the **Pearson chi-squared** is $X^2 = \sum e_{ij}^2$.

Deviance residual is defined as:

$$\sqrt{d_i} \times \operatorname{sign}(y_i - \hat{\mu}_i)$$

Where $d_i = 2\omega_i \left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right]$, the **likelihood ratio** statistic for testing independence is $G^2 = \sum d_{ij}$.

Standardized residual is defined as :

$$r_i = \frac{y_i - \hat{\mu}_i}{\left\{\left[\operatorname{var}(Y_i)\right]\left(1 - \hat{h}_i\right)\right\}^{1/2}} = \frac{e_i}{\sqrt{1 - \hat{h}_i}}$$

Then, the **Wald chi-squared** statistic is $Z_i^2 = \sum r_{ij}^2$.

- For model M_3 By applying the sequel as followed, three statistics are calculated:

```
#LR
deviance(m3)
#Pearson X2
pearson.resid2 <- resid(m3, type="pearson") # Pearson residuals
sum(pearson.resid2^2)
#Wald
sum(coef(summary(m3))[,"z value"])
```

Which by calculation, $G^2 = 94.84139$, $X^2 = 88.49485$, $Z^2 = 14.017$, they both asymptotically follows normal distribution for large samples. For small sample size, likelihood ratio G^2 is more reliable than Wald statistic Z^2. However, all of them larger than the chi-squared statistic of 0.95 quantile and 3 degrees of freedom which is 7.914728. We should reject the null hypothesis, and conclude that our model **does not have good fit** as the **saturated models**.

- For model M_4, apply the same process, it has the same G^2 and X^2 statistic, it makes sense since the two models only differs in the way of encoding, the Wald chi-squared statistic $Z^2 = 20.72486$, which also yields that the model does not have good fit. as the saturated model

```
#LR G2
deviance(m4)
#Pearson X2
pearson.resid2 <- resid(m4, type="pearson") # Pearson residuals
sum(pearson.resid2^2)
#Wald X2
sum(coef(summary(m4))[,"z value"])
qchisq(0.95,3)
```

- If consider the testing **joint significance of all predictors**, compare the main effect model with the **null model**.

  By similar process, all of the statistics could be calculated, ANOVA table would also be able to present the result. The ANOVA result is provided, below is a simple illustration from model M_1 and M_3.

```
Anova(m1, type = 2)
```

```
Analysis of Deviance Table (Type II tests)

Response: as.matrix(Tricks)
                               LR Chisq Df Pr(>Chisq)
as.matrix(dummy_r)               21.000  3  0.0001053 ***
Tricks.Card                      46.808  1  7.829e-12 ***
as.matrix(dummy_r):Tricks.Card   94.841  3  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compared with the null model, model M_3 contains the main effect terms, and has $G^2 = 67.808$, which is larger than the chi square with degrees of freedom 4 and 0.95 quantile (9.487729), reject the hypothesis implying model M_3 has significantly **better fit** than the **null model**. Similar concllusions are also made to model M_4

(f) Calculate both the likelihood-ratio and Wald chi-square statistic on the interaction effect, and comment on the adequacy of the sample size.

For studying the interaction effect, concluded by the difference of saturated models (M_1, M_2) and the main-effects models (M_3, M_4). Three statistics: likelihood-ratio (G^2), Wald chi-square (Z^2) and Pearson chi-squared (X^2) is calculated as followed.

```
#LR G2
deviance(m3)-deviance(m1)
deviance(m4)-deviance(m2)
anova(m1)
#Pearson
pearson.resid2 <- resid(m3, type="pearson") # Pearson residuals
sum(pearson.resid2^2)
pearson.resid3 <- resid(m4, type="pearson") # Pearson residuals
sum(pearson.resid3^2)
#Wald X2
sum(coef(summary(m1))[,"z value"][6:8]^2)
sum(coef(summary(m2))[,"z value"][6:8]^2)
(coef(summary(m1))[,"z value"][6:8]^2)
```

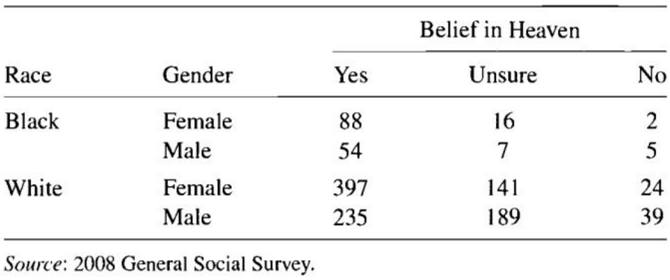
The table below provide the result of studying the interaction effects for both ways of categorical variable encoding. The significant values are in bold.

	X^2	G^2	Z^2
interactions effect (reference coding)	**94.84139**	**88.49485**	46.66357
interactions effect (effect coding)	**94.84139**	**88.49485**	$8.835438 * 10^{-7}$

Inferences are made, since most of the statistics are larger than chi squared statistics with 3 degrees of freedom, $x^2_{3,0.95} = 7.814728$ (except for Z^2) . This indicates that interactions effect are **very important** for model fitting. And thus, providing significantly better fitting when included.

When the **sample size** is small, the likelihood ratio statistic has more reliable results than Wald statistic (Z^2), in this problem, the likelihood ratio statistic always provide a stronger evidence than Wald. The evidently insignificant terms is the intersection of magician Green and trick type ($z^2_{gc} = 2.55 * 10^{-7}$), which is probably caused by the inadequacy of the sample size that we only collect the successful result.

2. For the table below, let Y = belief in existence of heaven, x1 = gender (1 = females, 0 =males), and x2 = race (1 = blacks, 0 = whites).



		Belief in Heaven		
Race	Gender	Yes	Unsure	No
Black	Female	88	16	2
	Male	54	7	5
White	Female	397	141	24
	Male	235	189	39

Source: 2008 General Social Survey.

(a) Fit a baseline-category logit model with main effects only to the data and interpret the conditional gender and race effects respectively.

- Let $\pi_j(x) = P(Y = j \mid x)$ at a fixed setting x for explanatory variables, with $\sum_j \pi_j(x) = 1$, it is able to treat the counts at the J categories of Y as multinomial with probabilities $\{\pi_1(x), \cdots, \pi_J(x)\}$. The baseline logit model is constructed by pairing each response with a baseline category:

$$\log \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}_j' \mathbf{x}, \quad j = 1, \ldots, J-1$$

In this problem, we have **Y** as a multi-category variable of three different believes in heaven, **x** be predictors of two variables: race and gender. Then the model under this setting could be written as:

$$\begin{cases} \log \dfrac{\pi_1(x)}{\pi_3(x)} = \alpha_1 + \beta_{11}x_1 + \beta_{21}x_2 \\ \log \dfrac{\pi_2(x)}{\pi_3(x)} = \alpha_2 + \beta_{21}x_1 + \beta_{22}x_2 \end{cases}$$

This model simultaneously describes the effects of $\{x_i\}$ on these $J-1$ logits.

- Using the sequel as followed, where the belief in heaven noted as "No" is set as the baseline. For race variable x_1, black is set to be 0, and white is 1; for gender variable x_2, female is set to be 0, and male is 1.

```
race=c(0,0,1,1)
gender=c(0,1,0,1)
Yes=c(88,54,397,235)
Unsure=c(16,7,141,189)
No=c(2,5,24,39)
belief=cbind(race,gender,Yes,Unsure,No)
belief_df=as.data.frame(belief)
l1=vglm(formula = cbind(Yes,Unsure,No) ~ race + gender,family=multinomial,
data=belief_df)
summary(l1)
```

```
Call:
vglm(formula = cbind(Yes, Unsure, No) ~ race + gender, family = multinomial,
    data = belief_df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1   3.5009     0.4179   8.376  < 2e-16 ***
(Intercept):2   1.3638     0.4602   2.964  0.00304 **
race:1         -0.6727     0.4114  -1.635  0.10205
race:2          0.4757     0.4533   1.049  0.29397
gender:1       -1.0339     0.2587  -3.997 6.41e-05 ***
gender:2       -0.3087     0.2697  -1.145  0.25235
---
```

Thus, the **baseline logistic model** L_1 with main effect could be written as:

$$\begin{cases} \log \dfrac{\pi_1(x)}{\pi_3(x)} = 3.5009 - 0.6727x_1 - 1.0339x_2 \\ \log \dfrac{\pi_2(x)}{\pi_3(x)} = 1.3638 + 0.4757x_1 - 0.3087x_2 \end{cases}$$

- Interpretation of the **conditional gender and race effect** is made based on the estimated value of conditional probability (β_i) combined with the coefficients estimated:

  When the race is fixed, the probability of a female believe in heaven is larger than male (Yes), but the probability of a "Unsure" and "No" is less than male. When the gender is fixed, the probability of a black people believe in heaven is larger than white (Yes), but the probability of a "Unsure" and "No" is less than white.

  ```
  predict(l1,type="response")
  ```

Race	Gender	Yes	Unsure	No
Black	Female	0.8709519	0.1027716	0.02627655
Black	Male	0.7527137	0.1834274	0.06385887
White	Female	0.6987173	0.2599755	0.04130727
White	Male	0.5168918	0.3971788	0.08592941

(b) Comment on the goodness of fit. Conduct a likelihood-ratio test of whether opinion is independent of gender, given race.

- First carry out the **goodness of fit** test by the **Likelihood ratio**
  - Compare with the null model, $G^2 = 72.741$ which is greater than $x^2_{4,0.95} = 9.488$. Reject the null hypothesis, model L_1 has good fit and is useful.

    ```
    lrtest(l1)
    ```

    ```
    Model 1: cbind(Yes, Unsure, No) ~ race + gender
    Model 2: cbind(Yes, Unsure, No) ~ 1
      #Df  LogLik Df  Chisq Pr(>Chisq)
    1   2 -21.792
    2   6 -58.162  4 72.741  5.986e-15 ***
    ```

  - Compare with the saturated model, $G^2 = 6.0748$ which is greater than $x^2_{2,0.95} = 5.991$. Also reject the null hypothesis model L_1 does not have a good fit as the saturated model, though it is very close.

```
l0=vglm(formula = cbind(Yes,Unsure,No) ~ gender*race,family=multinomial,
data=belief_df)
lrtest(l1,l0)
```

```
Likelihood ratio test

Model 1: cbind(Yes, Unsure, No) ~ race + gender
Model 2: cbind(Yes, Unsure, No) ~ gender * race
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -21.792
2   0 -18.754 -2 6.0748    0.04796 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Then conduct a **likelihood-ratio test** of checking whether opinion is **independent** of gender given race.

Fix the race to be black or white, I separately conducted the likelihood ratio test to the group data to find if there is any independence. The Likelihood ratio statistic is thus defined as:

$$G^2 = 2\sum_i\sum_j n_{ij}\log(n_{ij}/\mu_{ij})$$

With the degrees of freedom equals $(I-1)(J-1)$. Using the sequel:

```
black_b = cbind(Yes[1:2],Unsure[1:2],No[1:2])
white_b = cbind(Yes[3:4],Unsure[3:4],No[3:4])
#black
n_bs = apply(black_b,1,sum)
n_bs2 = apply(black_b,2,sum)
n_bt = sum(black_b)
mu_b = (n_bs %*% t(n_bs2))/n_bt
(Gb = 2*sum(black_b*log(black_b/mu_b)))
#white
n_ws = apply(white_b,1,sum)
n_ws2 = apply(white_b,2,sum)
n_wt = sum(white_b)
mu_w = (n_ws %*% t(n_ws2))/n_wt
(Gw = 2*sum(white_b*log(white_b/mu_w)))
```

For each subtable, we calculate the likelihood ratio statistic. Then, the $G^2_{black} = 3.7783$, while $G^2_{white} = 43.0282$. Since they both follows asymptotically chi-squared distribution with 2 degrees of freedom, then $x^2_{2,0.95} = 5.9915$. For the black people, do not reject null hypothesis, opinions are independent of gender. However, for white people, we should reject the null hypothesis, thus opinions are not independent of gender. Indeed, when gender is fixed, it could be seen as a constant. We could carry out the likelihood ratio test for model L_1 compared with model only on race L_{11}.

```
l0=vglm(formula = cbind(Yes,Unsure,No) ~ race,family=multinomial,
data=belief_df)
#check independence
deviance(l0)-deviance(l1)
```

Then $G^2 = 40.73175$, which is larger than $x^2_{0.95} = 5.9914$, thus should reject the hypothesis. Opinions about belief depends on gender given race.

(c) Treating belief in heaven as ordinal, fit and interpret: (i) a cumulative logit model and (ii) a cumulative probit model. Compare the results and state interpretations in each case.

In this problem, belief is treated as an ordinal variable. Then, a cumulative logit model and a cumulative probit model are derived and compared.

- The **cumulative logit model with proportional odds** is defined as:

  $$\text{logit}[P(Y \leq j \mid \mathbf{x})] \mid = \alpha_j + \boldsymbol{\beta}'\mathbf{x}$$

  With the logit link, setting **parallel = TRUE** will fit a proportional odds model. In practice, the validity of the proportional odds assumption needs to be checked, e.g., by a likelihood ratio test (LRT). If acceptable on the data, then numerical problems are less likely to occur during the fitting, and there are less parameters. While by default, the non-parallel cumulative logit model is fitted,

  $$\eta_j = \text{logit}(P[Y \leq j])$$

  where $j = 1, 2, \cdots, M$ and η_j are not constarined to be parallel.

- First try a **cumulative model** (L_2) with proportional odds ratio:

```
l2 <- vglm(cbind(Yes,Unsure,No) ~ race + gender,
            family=cumulative(parallel=TRUE), data=belief_df)
summary(l2)
```

```
Call:
vglm(formula = cbind(Yes, Unsure, No) ~ race + gender, family =
cumulative(parallel = TRUE),
    data = belief_df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1   1.8623     0.2096   8.884  < 2e-16 ***
(Intercept):2   4.1084     0.2422  16.966  < 2e-16 ***
race           -1.0165     0.2106  -4.827 1.39e-06 ***
gender         -0.7696     0.1225  -6.281 3.37e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Exponentiated coefficients:
     race    gender
0.3618767 0.4632146
```

The model provided could be notes as:

$$\begin{cases} \text{logit}[P(Y \leq 1 \mid \mathbf{x})] = 1.8623 - 1.0165x_1 - 0.7696x_2 \\ \text{logit}[P(Y \leq 2 \mid \mathbf{x})] = 4.1084 - 1.0165x_1 - 0.7696x_2 \end{cases}$$

With 3 response categories, the model has two intercepts.

To check the validity of the proportional odds assumption, use the cumulative logit model (L_{22}) not having the proportional odds assumption to see if we can get a better fit. The likelihood ratio test indicates it does no provide a evidently better fit ($G^2 = 2.831981$, $p = 0.2426851$). Thus, we use model L_2 since it has less estimation of coefficients.

```
l2_2<- vglm(cbind(Yes,Unsure,No) ~ race + gender, family=cumulative,
data=belief_df)
pchisq(deviance(l2)-deviance(l2_2),df=df.residual(l2)-
df.residual(l2_2),lower.tail=FALSE)
```

```
Call:
vglm(formula = cbind(Yes, Unsure, No) ~ race + gender, family = cumulative,
    data = belief_df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1   1.8864     0.2119   8.903  < 2e-16 ***
(Intercept):2   3.5196     0.4079   8.629  < 2e-16 ***
race:1         -1.0462     0.2128  -4.917 8.78e-07 ***
race:2         -0.3387     0.4027  -0.841  0.40033
gender:1       -0.7682     0.1244  -6.176 6.55e-10 ***
gender:2       -0.8302     0.2555  -3.249  0.00116 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Exponentiated coefficients:
   race:1    race:2  gender:1  gender:2
0.3512583 0.7127025 0.4638615 0.4359809
```

- Then, conduct a **cumulative probit model** (L_3):

  Instead of using the link function of logit, then consider the model:

  $$\Phi^{-1}[P(Y \leq j)] = \alpha_j - \beta x$$

  Similarly, use the parallel assumption reduce the calculation complexity.

```
l3 <- vglm(cbind(Yes,Unsure,No) ~ race + gender,
                   family=cumulative(link=probit, parallel=TRUE), belief_df)
summary(l3)
```

```
Call:
vglm(formula = cbind(Yes, Unsure, No) ~ race + gender, family =
cumulative(link = probit,
    parallel = TRUE), data = belief_df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  1.06083    0.11354   9.343  < 2e-16 ***
(Intercept):2  2.29599    0.12661  18.134  < 2e-16 ***
race          -0.54371    0.11520  -4.720 2.36e-06 ***
gender        -0.44936    0.07201  -6.241 4.36e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Exponentiated coefficients:
     race    gender
0.5805920 0.6380384
```

The model could be further written as:

$$\begin{cases} \Phi(Y \leq 1 \mid \mathbf{x}) = 1.06083 - 0.54371x_1 - 0.44936x_2 \\ \Phi(Y \leq 2 \mid \mathbf{x}) = 2.29599 - 0.54371x_1 - 0.44936x_2 \end{cases}$$

- Then, we try to **compare** the result and give some **interpretations**.
  - To **compare** the two different types of logistic models, use pseudo R^2 which is defined as:

$$R^2_{pseudo} = 1 - \frac{\text{Residual Deviance}}{\text{Null Deviance}}$$

```
#Null model
l2_0<-vglm(cbind(Yes,Unsure,No) ~ 1,
           family=cumulative(parallel=TRUE), data=belief_df)
l3_0<-vglm(cbind(Yes,Unsure,No) ~ 1,
           family=cumulative(link=probit, parallel=TRUE), belief_df)
#null deviance
null_dl2=l2_0@criterion$deviance
null_dl3=l3_0@criterion$deviance
#deviance
dl2=l2@criterion$deviance
dl3=l3@criterion$deviance
#psuedo R2
1-dl2/null_dl2
1-dl3/null_dl3
```

For model L_2, $R^2_{pseudo} = 0.8826$; for model L_3, $R^2_{pseudo} = 0.8336$. Which possibly indicates that the Logistic basline model has a better fitting.

  - **Interpretations**: both of the models indicate that white and male people tends to be less likely to believe in heaven or unsure about it. They are more of not believe in heaven than other opinions. Both models provide practical result for this problem.

(d) (*Optional)* Use Bayesian methods to fit the model of (a) with uninformative priors. Interpret the results and compare them with the ML estimates.

- In this problem, we are interested in conditional probability model.

$$P\left(y = k \mid \boldsymbol{\beta}, \mathbf{x}_i\right) = \operatorname{logit}\left(\boldsymbol{\beta}^T \mathbf{x}_i\right) = \operatorname{logit}\left(\sum_j \beta_j x_{i,j}\right)$$

From the *zelig* project, it uses Bayesian multinomial logistic regression to model unordered categorical variables that sets the baseline to the first dependent variable of alphabetic order. The model is estimated via a random walk Metropolis algorithm or a slice sampler. By default, it leads to an **improper prior** ($\left[E\left(\partial^2 \log f(y \mid \theta)/\partial\theta^2\right)\right]^{1/2}$ with a single θ) as requested.

By applying the sequel as followed, the empirical mean and standard deviation, as well as quantiles (confidence interval) of the multinomial logistic bayesian model (L_4) could be derived.

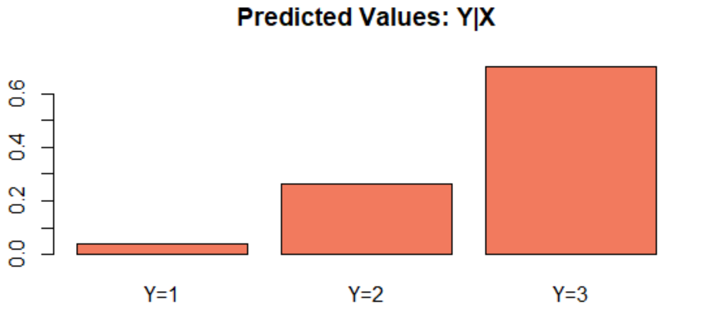
```
Belief1<-factor(c("Yes","Unsure","No"),levels=c("Yes","Unsure","No"))
Gender1<-factor(c("Female","Male"),levels=c("Female","Male"))
Race1<-factor(c("Black", "White"),levels=c("Black", "White"))
Data2<-expand.grid(Belief=Belief1,Gender=Gender1,Race=Race1)
People<-c(88,16,2,54,7,5,397,141,24,235,189,39)
Data3<-structure(.Data=Data2[rep(1:nrow(Data2),People),],row.names=1:1197)
z.out <- zelig(Belief ~ Gender + Race, model = "mlogit.bayes", data = Data3)
x.out <- setx(z.out)
z.out$geweke.diag()
z.out$heidel.diag()
z.out$raftery.diag()
summary(z.out)
```

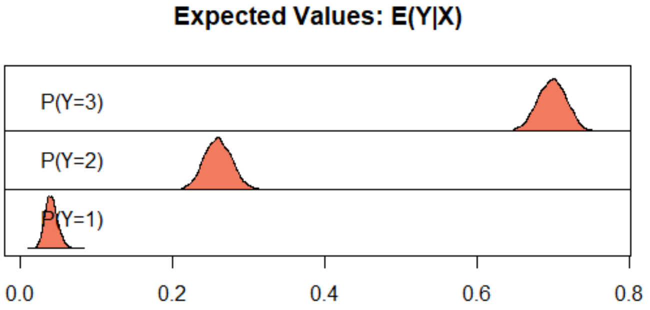
The grouped data is divided into ungrouped data in this problem.

```
Model:

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                       Mean     SD Naive SE Time-series SE
(Intercept).Unsure   1.4247 0.4720 0.004720       0.006353
(Intercept).Yes      3.5846 0.4314 0.004314       0.006360
GenderMale.Unsure   -0.3143 0.2745 0.002745       0.003269
GenderMale.Yes      -1.0428 0.2648 0.002648       0.003134
RaceWhite.Unsure     0.4307 0.4667 0.004667       0.006494
RaceWhite.Yes       -0.7393 0.4269 0.004269       0.006185

2. Quantiles for each variable:

                       2.5%     25%     50%     75%    97.5%
(Intercept).Unsure   0.5387  1.1076  1.4146  1.7346  2.37963
(Intercept).Yes      2.7919  3.2926  3.5601  3.8571  4.48944
GenderMale.Unsure   -0.8479 -0.4984 -0.3143 -0.1307  0.22606
GenderMale.Yes      -1.5642 -1.2189 -1.0394 -0.8636 -0.54224
RaceWhite.Unsure    -0.5109  0.1317  0.4407  0.7527  1.31060
RaceWhite.Yes       -1.6469 -1.0107 -0.7194 -0.4453  0.04185
```

The result also indicates that Male and white people has less likely to belief in Heaven. The confidence interval could be acquired by using the 2.5% and 97.5% quantiles. Output the simulated expected values(probabilities) of each of the J categories given the specified values of x.

```
s.out <- sim(z.out, x = x.out)
plot(s.out)
```





Consider when the predictor is the largest population(female, white), simulate quantities of interest from the posterior distribution predicted values ($P(Y = 1 \mid x) = 0.04139113$, $P(Y = 2 \mid x) = 0.26012694$, $P(Y = 3 \mid x) = 0.69848193$) and expected probability of each outcome. **Compared with estimation in L_1, L_2 and L_3** as in the table, this is the closest ($MSE = 0.01224$) to the observed value so far.

	Yes	Unsure	No	MSE
L_1	0.6987173	0.2599755	0.04130727	1.2110
L_2	0.6997019	0.2568573	0.04344078	1.2278
L_3	0.6974635	0.2626739	0.03986261	1.2164
L_4	0.69848193	0.26012694	0.04139113	**0.0122**
observed	0.70640569	0.25088968	0.04270463	0