Final Project

Tujuan:
Tugas ini bertujuan untuk memahami lebih dalam dan menerapkan beberapa metode pengolahan data kategorik untuk memperoleh insights dan menganalisis hubungan antar pengukuran yang bersifat kategorik.

Cakupan tugas:

1. Pemahaman konstekstual, berupa interpretasi kuantitas yang dipelajari dalam konteks contoh permasalahan nyata
2. Pengolahan data, analisis dan interpretasi hasil

Due date: Jumat 30 Desember 2022 pukul 08.00 WIB

Anggota kelompok:

No	Nama	NPM	Kontribusi	Tingkat kontribusi
1	Andini Assyahidah	2006571040	Terlibat aktif diskusi, menuliskan ringkasan hasil diskusi, mengerjakan beberapa bagian 2 dan bagian 3	100%
2	Annisa Fairuz Zahira	2006571015	Terlibat aktif diskusi	100%
3	Auranissa Efrida Putri	2006571192	Terlibat aktif diskusi, menuliskan ringkasan hasil diskusi, mengerjakan bagian 1, beberapa bagian 2, dan bagian 3	100%
4	Laily Nur Azizah	2006464234	Terlibat aktif diskusi, menuliskan ringkasan hasil diskusi, mengerjakan bagian 1, beberapa bagian 2	100%
5	Nadhila Nur Qamarina	2006521742	Terlibat aktif diskusi	100%

Instruksi:
Gunakan data berikut untuk analisis seperti panduan di bawah ini.

Safety Equipment in Use	Whether Ejected	Injury	
		Nonfatal	Fatal
Seat belt	Yes	1,105	14
	No	411,111	483
None	Yes	4,624	497
	No	157,342	1,008

Source: Florida Department of Highway Safety and Motor Vehicles.

Bagian 1. Pendahuluan

[**C4, 10 points**] Berdasarkan data pengukuran di atas, tuliskan beberapa insights (atau hipotesis) yang mungkin bisa diperoleh, atau dikonfirmasi, dari analisis yang dapat dilakukan pada data tersebut.]

- Dapat mencari tahu adanya asosiasi penggunaan seatbelt terhadap Injury.
- Dapat mencari tahu adanya asosiasi saat penumpang terpental terhadap Injury.

- Dapat mencari tahu adanya asosiasi saat penggunaan seatbelt terhadap terpentalnya penumpang atau tidak.
- Dapat mencari tahu tingkat resiko perilaku menggunakan seatbelt terhadap Injury.
- Dapat mencari tahu tingkat resiko perilaku menggunakan seatbelt terhadap terpentalnya penumpang atau tidak.
- Dapat mencari tahu tingkat resiko penumpang terpental terhadap Injury.
- Menggunakan model log-lin untuk melihat hubungan antara ketiga variabel kategorik.
- Dapat mencari model yang terbaik untuk digunakan.
- Mengetahui kelebihan/kekurangan model Log Linear dan model logistik

Bagian 2. Lakukan pengolahan dan analisis data dengan panduan dari pertanyaan-pertanyaan berikut.

Safety Use Equipment	Whether Ejected	Injury	
		Nonfatal	fatal
seatbelt	**yes**	1105	14
	no	411111	483
none	**yes**	4624	497
	no	157342	1008

1. [**C4, 10 points**] Apakah ada asosiasi antara penggunaan seat belt dengan fatal/non-fatal injury yang dialami penumpang kendaraan tersebut jika terjadi kecelakaan? Jelaskan prosedur pengolahan data yang dilakukan untuk pengecekan asosiasi ini.

Safety Use Equipment	Injury	
	Non Fatal	Fatal
Seat Belt	412216	497
None	161966	1505

Untuk melihat pengaruh seatbelt terhadap injury, dapat dihitung perbandingan odds antara memakai seatbelt dengan yang tidak

- Odds memakai seatbelt = $O_1 = \frac{412216}{497} = 829.41$
- Odds tidak memakai seatbelt = $O_2 = \frac{161966}{1505} = 107.62$
- Odds rasio = $\widehat{\theta} = \frac{O_1}{O_2} = \frac{829.41}{107.62} = 7.71$ -> terlihat sepintas apabila kita lihat dari odds rasionya, orang-orang yang **memakai seatbelt cenderung** mempunyai resiko yang lebih besar untuk terkena **injury yang non fatal** dibandingkan mereka yang tidak memakai seatbelt, yang berarti dapat dinyatakan ada indikasi hubungan (asosiasi) antara penggunaan seatbelt dengan terkenanya injury fatal/nonfatal. Akan tetapi, untuk mendapatkan kesimpulan yang lebih kuat secara statistika, perlu dilakukan pengujian hipotesis, atau ekivalen dengan melakukan perhitungan interval kepercayaan dari rasio odds.
- log $\widehat{\theta} = log(7.71) = 0.8871$
- S.E log $\widehat{\theta} = \sqrt{\frac{1}{412216} + \frac{1}{497} + \frac{1}{161966} + \frac{1}{1505}} = 0.05182$
- Maka, interval kepercayaan 95% untuk log θ

  $$0.8871 \pm 1.96(0.05182) \approx (0.7855, 0.9886)$$
- Atau dalam θ, (exp(0.7855), exp(0.9886)) = **(2.1935, 2.6875)**

- Berdasarkan hasil di atas, karena interval kepercayaan untuk θ tidak memuat nilai 1 (yang menyatakan peluang bahwa rasio odds akan bernilai 1; a.k.a kedua odds sama adalah 0.95); maka dapat disimpulkan bahwa **terdapat bukti yang kuat** secara statistika pada taraf signifikansi 5% **terdapat asosiasi** antara penggunaan seatbelt terhadap fatal/non fatal injury pada penumpang kecelakaan.

2. [**C4, 10 points**] Pada kondisi dimana terjadi "Ejected", apakah penggunaan seat belt mengurangi resiko terjadinya fatal injury pada penumpang (jika terjadi kecelakaan)?

Safety Use Equipment	**Whether Ejected**	**Injury**	
		Nonfatal	**Fatal**
seatbelt	**yes**	1105	14
none	**yes**	4624	497

- $\pi_1 = \frac{1105}{1119} = 0.9875,\ 1 - \pi_1 = 0.0125$
- $\pi_2 = \frac{4624}{5121} = 0.9029,\ 1 - \pi_2 = 0.0971$
- $r = \frac{1-\pi_1}{1-\pi_2} = 0.1287332647 < 1$. Diduga ada perbedaan resiko yang berarti antara kedua kelompok ini. Akan tetapi, untuk mendapatkan kesimpulan lebih kuat secara statistika, akan dihitung interval kepercayaan dari resiko relative tersebut.
- $\sigma(log\ r) = \sqrt{\frac{0.9875}{14} + \frac{0.9029}{497}} = 0.26898$
- $log\ r = log(0.1287) = -0.89042$

- Maka confidence interval untuk r dengan 95%;

$$-0.8904 \pm 1.96(0.26898) = (exp(-1.417), exp(-0.3632))$$
$$= \textbf{(0.2423, 0.6954)}$$

- Kesimpulan:
  - ADA perbedaan tingkat resiko menggunakan seatbelt dan tidak terhadap fatal injury pada kondisi ejected.
  - Tingkat resiko fatal injury yang memakai seatbelt berada pada rentang (0.2423, 0.6954) dari orang yang tidak memakai seatbelt. Hal ini menunjukan penggunaan seat belt mengurangi resiko terjadinya fatal injury pada penumpang. (karena tingkat resiko kecil)

3. [**C4, 10 points**] Apakah "whether ejected (yes/no)" merupakan confounding variabel? Jelaskan (tuliskan pengecekan apa yang dilakukan pada data untuk menentukan confounding atau tidak).

Menurut (e.g., .J Cornfield in 1954, as summarized by Greenhouse 2009), confounding variable harus mempunyai kekuatan asosiasi yang kuat antara confounding variable Z dan kedua variabel X dan Y. Akan dihitung masing-masing kekuatan asosiasi antara variabel XYZ(1) dan XYZ(2) dengan odds ratio.

Whether Ejected	**Safety Use Equipment**	**Injury**	
		Nonfatal	**Fatal**
Yes	**seatbelt**	1105	14
	none	4624	497

- $\theta_{XY(1)} = \frac{\frac{1105}{5729} X \frac{497}{511}}{\frac{14}{511} X \frac{4624}{5729}} = 8.483$
- $\log \widehat{\theta} = log(9.483) = 0.9286$
- S.E $\log \widehat{\theta} = \sqrt{\frac{1}{1105} + \frac{1}{497} + \frac{1}{14} + \frac{1}{4624}} = 0.2731$
- Maka, interval kepercayaan 95% untuk $\log \theta$

$$0.9286 \pm 1.96(0.2731) \approx (0.393324, 1.321924)$$

- Atau dalam θ, (exp(0.393324), exp(1.321924)) = **(1.481898447, 3.750630653)**
- Berdasarkan hasil di atas, karena interval kepercayaan untuk θ memuat nilai 1 (yang menyatakan peluang bahwa rasio odds akan bernilai 1; a.k.a kedua odds sama adalah 0.95); maka dapat disimpulkan bahwa **tidak terdapat bukti yang kuat** secara statistika pada taraf signifkansi 5% terdapat asosiasi antara penggunaan seatbelt, fatal/non fatal injury terhadap ejected/ no pada penumpang kecelakaan.

Whether Ejected	**Safety Use Equipment**	**Injury**	
		Nonfatal	**fatal**
No	**seatbelt**	411111	483
	none	157342	1008

- $\theta_{XY(2)} = \frac{\frac{1105}{5729} X \frac{497}{511}}{\frac{14}{511} X \frac{4624}{5729}} = 5.4529$
- $\log \widehat{\theta} = log(9.483) = 0.7366$
- S.E $\log \widehat{\theta} = \sqrt{\frac{1}{41111} + \frac{1}{483} + \frac{1}{1008} + \frac{1}{157342}} = 0.05542$
- Maka, interval kepercayaan 95% untuk $\log \theta$

$$0.7366 \pm 1.96(0.05542) \approx (0.6279768, 0.8452232)$$

- Atau dalam θ, (exp(0.393324), exp(1.321924)) = **(1.8738, 2.3285)**
- Berdasarkan hasil di atas, karena interval kepercayaan untuk θ memuat nilai 1 (yang menyatakan peluang bahwa rasio odds akan bernilai 1; a.k.a kedua odds sama adalah 0.95); maka dapat disimpulkan bahwa **tidak terdapat bukti yang kuat** secara statistika pada taraf signifkansi 5% **terdapat asosiasi** antara penggunaan seatbelt, fatal/non fatal injury terhadap ejected/ no pada penumpang kecelakaan.

Maka, dapat dikatakan ejected(yes/no) bukan merupakan suatu confounding variable.

4. [**C4, 20 points**] Tentukan model terbaik yang menjelaskan keterkaitan antara penggunaan seat belt, ejected atau tidak, dan fatal atau tidaknya injury berdasarkan data di atas. Tuliskan prosedurnya dengan runut dan jelas sehingga diperoleh model terbaik tersebut (jika ada pengujian hipotesis, tuliskan prosedurnya). Tuliskan model yang menjadi acuannya (saturated model).

   Untuk melihat keterkaitan antara penggunaan seat belt, ejected atau tidak, dan fatal atau tidaknya injury akan digunakan model log linear. Pertama, akan dimodelkan Freq sebagai fungsi dari tiga variabel tersebut dengan menggunakan fungsi glm. Kemudian, akan digunakan argumen family poisson karena menghitung jumlah pemodelan pada data, sehingga diasumsikan ketiga variabel tersebut saling independen.

- Model 1

```
> mod0 <- glm(Freq ~ Ejected + Injury + Safety,
+             data = safety.df, family = poisson)
> summary(mod0)

Call:
glm(formula = Freq ~ Ejected + Injury + Safety, family = poisson,
    data = safety.df)

Deviance Residuals:
      1       2       3       4       5       6       7       8
-60.146   6.708  -0.395 -28.814  56.492  -9.496  58.171  16.913

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.401580   0.012687   662.2   <2e-16 ***
Ejectedno    4.514558   0.012728   354.7   <2e-16 ***
Injuryfatal -5.658800   0.022388  -252.8   <2e-16 ***
Safetynone  -0.926117   0.002922  -316.9   <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: 1624865  on 7  degrees of freedom
Residual deviance:   11444  on 4  degrees of freedom
AIC: 11529

Number of Fisher Scoring iterations: 6
```

```
> cbind(mod0$data, fitted(mod0))
  Ejected   Injury   Safety   Freq fitted(mod0)
1     yes nonfatal seatbelt   1105 4.454100e+03
2      no nonfatal seatbelt 411111 4.068249e+05
3     yes    fatal seatbelt     14 1.553011e+01
4      no    fatal seatbelt    483 1.418476e+03
5     yes nonfatal     none   4624 1.764219e+03
6      no nonfatal     none 157342 1.611388e+05
7     yes    fatal     none    497 6.151301e+00
8      no    fatal     none   1008 5.618425e+02
```

Pada model pertama dapat dilihat bahwa koefisien estimasinya cukup signifikan dan memiliki nilai p-value yang mendekati 0, serta memiliki nilai AIC sebesar 11529 dan residual deviance sebesar 11444.

- Model 2

```
> mod1 <- glm(Freq ~ (Ejected + Injury + Safety)^2,
+             data = safety.df, family = poisson)
> summary(mod1)

Call:
glm(formula = Freq ~ (Ejected + Injury + Safety)^2, family = poisson,
    data = safety.df)

Deviance Residuals:
      1       2       3       4       5       6       7       8
 0.20704  -0.01071  -1.59987   0.31400  -0.10095   0.01731   0.30951  -0.21583

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              7.00137    0.02992  233.99   <2e-16 ***
Ejectedno                5.92527    0.02996  197.76   <2e-16 ***
Injuryfatal             -3.96315    0.06944  -57.07   <2e-16 ***
Safetynone               1.43913    0.03321   43.33   <2e-16 ***
Ejectedno:Injuryfatal   -2.79779    0.05526  -50.63   <2e-16 ***
Ejectedno:Safetynone    -2.39964    0.03334  -71.97   <2e-16 ***
Injuryfatal:Safetynone   1.71732    0.05402   31.79   <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: 1.6249e+06  on 7  degrees of freedom
Residual deviance: 2.8540e+00  on 1  degrees of freedom
AIC: 93.853

Number of Fisher Scoring iterations: 3
```

```
> cbind(mod1$data, fitted(mod1))
  Ejected   Injury   Safety   Freq fitted(mod1)
1     yes nonfatal seatbelt   1105   1098.13193
2      no nonfatal seatbelt 411111 411117.86807
3     yes    fatal seatbelt     14      20.86807
4      no    fatal seatbelt    483     476.13193
5     yes nonfatal     none   4624    4630.86807
6      no nonfatal     none 157342 157335.13193
7     yes    fatal     none    497     490.13193
8      no    fatal     none   1008    1014.86807
```

Pada model kedua dapat dilihat bahwa koefisien estimasinya cukup signifikan dan memiliki nilai p-value yang mendekati 0, serta memiliki nilai AIC sebesar 93.853 dan residual deviance sebesar 2.854.

- Model 3

```
> mod2 <- glm(Freq ~ Ejected * Injury * Safety,
+             data = safety.df, family = poisson)
> summary(mod2)

Call:
glm(formula = Freq ~ Ejected * Injury * Safety, family = poisson,
    data = safety.df)

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

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                       7.00760    0.03008 232.943  < 2e-16 ***
Ejectedno                         5.91902    0.03012 196.493  < 2e-16 ***
Injuryfatal                      -4.36854    0.26895 -16.243  < 2e-16 ***
Safetynone                        1.43141    0.03348  42.748  < 2e-16 ***
Ejectedno:Injuryfatal            -2.37806    0.27278  -8.718  < 2e-16 ***
Ejectedno:Safetynone             -2.39186    0.03362 -71.153  < 2e-16 ***
Injuryfatal:Safetynone            2.13812    0.27306   7.830 4.87e-15 ***
Ejectedno:Injuryfatal:Safetynone -0.44197    0.27863  -1.586    0.113
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  1.6249e+06  on 7  degrees of freedom
Residual deviance: -5.0009e-11  on 0  degrees of freedom
AIC: 92.999

Number of Fisher Scoring iterations: 3
```

```
> mod2 <- glm(Freq ~ Ejected * Safety * Injury,
+             data = safety.df, family = poisson)
> cbind(mod2$data, fitted(mod2))
  Ejected   Injury   Safety   Freq fitted(mod2)
1     yes nonfatal seatbelt   1105         1105
2      no nonfatal seatbelt 411111       411111
3     yes    fatal seatbelt     14           14
4      no    fatal seatbelt    483          483
5     yes nonfatal     none   4624         4624
6      no nonfatal     none 157342       157342
7     yes    fatal     none    497          497
8      no    fatal     none   1008         1008
```

Pada model ketiga dapat dilihat bahwa koefisien estimasinya cukup signifikan dan memiliki nilai p-value yang mendekati 0, serta memiliki nilai AIC sebesar 92.999 dan residual deviance sebesar -5.009×10^{-11}.

- Model 4

```
> mod3 <- glm(Freq ~ (Safety*Injury)+(Ejected*Injury),
+             data = safety.df, family = poisson)
> summary(mod3)

Call:
glm(formula = Freq ~ (Safety * Injury) + (Ejected * Injury),
    family = poisson, data = safety.df)

Deviance Residuals:
      1       2       3       4       5       6       7       8
-55.779   4.703 -12.806   5.600  60.880  -7.535   5.506  -3.430

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            8.321897   0.013237  628.67   <2e-16 ***
Safetynone            -0.934161   0.002933 -318.55   <2e-16 ***
Injuryfatal           -3.478840   0.060372  -57.62   <2e-16 ***
Ejectedno              4.597378   0.013278  346.25   <2e-16 ***
Safetynone:Injuryfatal  2.042119   0.051818   39.41   <2e-16 ***
Injuryfatal:Ejectedno  -3.526545   0.052952  -66.60   <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: 1624865  on 7  degrees of freedom
Residual deviance:    7134  on 2  degrees of freedom
AIC: 7223

Number of Fisher Scoring iterations: 5
```

```
> mod3 <- glm(Freq ~ (Safety*Injury)+(Ejected*Injury),
+             data = safety.df, family = poisson)
> cbind(mod3$data, fitted(mod3))
  Ejected   Injury   Safety   Freq fitted(mod3)
1     yes nonfatal seatbelt   1105    4112.9563
2      no nonfatal seatbelt 411111  408103.0437
3     yes    fatal seatbelt     14     126.8566
4      no    fatal seatbelt    483     370.1434
5     yes nonfatal     none   4624    1616.0437
6      no nonfatal     none 157342  160349.9563
7     yes    fatal     none    497     384.1434
8      no    fatal     none   1008    1120.8566
```

Pada model keempat dapat dilihat bahwa koefisien estimasinya cukup signifikan dan memiliki nilai p-value yang mendekati 0, serta memiliki nilai AIC sebesar 7223 dan residual deviance sebesar 7134.

- Model 5

```
> mod4 <- glm(Freq ~ (Safety*Ejected)+Injury,
+             data = safety.df, family = poisson)
> summary(mod4)

Call:
glm(formula = Freq ~ (Safety * Ejected) + Injury, family = poisson,
    data = safety.df)

Deviance Residuals:
      1       2       3       4       5       6       7       8
 -0.303   1.478   3.956 -29.080  -6.817  -1.153  48.491  17.463

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           7.01671    0.02989  234.72   <2e-16 ***
Safetynone            1.52091    0.03300   46.09   <2e-16 ***
Ejectedno             5.90760    0.02993  197.35   <2e-16 ***
Injuryfatal          -5.65880    0.02239 -252.75   <2e-16 ***
Safetynone:Ejectedno -2.47614    0.03313  -74.74   <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: 1624865.3  on 7  degrees of freedom
Residual deviance:    3567.7  on 3  degrees of freedom
AIC: 3654.7

Number of Fisher Scoring iterations: 6
```

```
> cbind(mod4$data, fitted(mod4))
  Ejected   Injury   Safety   Freq fitted(mod4)
1     yes nonfatal seatbelt   1105   1115.11194
2      no nonfatal seatbelt 411111 410163.88186
3     yes    fatal seatbelt     14      3.88806
4      no    fatal seatbelt    483   1430.11814
5     yes nonfatal     none   4624   5103.20665
6      no nonfatal     none 157342 157799.79954
7     yes    fatal     none    497     17.79335
8      no    fatal     none   1008    550.20046
```

Pada model kelima dapat dilihat bahwa koefisien estimasinya cukup signifikan dan memiliki nilai p-value yang mendekati 0, serta memiliki nilai AIC sebesar 3654.7 dan residual deviance sebesar 3567.7.

Agar lebih mudah untuk membandingkan, akan dibuat tabel perbandingan *fitted values*,

Safety Equipment (S)	**Whether Ejected (E)**	**Injury (I)**	**Log Linear (fitted value)**				
			(S,E,I)	**(SE,I)**	**(SI,EI)**	**(SE,SI,EI)**	**(SEI)**
Seat Belt	Yes	Non Fatal	4454.1	1115.1	4113	1098.1	1105
		Fatal	15.1	3.9	126.9	20.9	14
	No	Non Fatal	406824.9	410163.9	408103	411117.9	411111
		Fatal	1418.5	1430.1	370.1	476.1	483
None	Yes	Non Fatal	1764.2	5103.2	1616	4630.9	4624
		Fatal	6.2	17.8	384.1	490.1	497
	No	Non Fatal	16113.9	157799.8	160350	157335.1	157342

		Fatal	561.8	550.2	1120.9	1014.9	1008

Untuk mendapatkan model loglinear yang cocok, akan dibandingkan fitted value dari model yang diuji. Dari hasil, didapatkan bahwa model SEI merupakan saturated model yang hasil perhitungannya sama dengan nilai observasi. Dari pengujian serta tabel perbandingan *fitted value*, diperoleh bahwa mod1 memiliki nilai yang **mendekati** dengan observasi data. Untuk menguatkan pilihan model yang dipilih, maka akan dilakukan pengujian Goodness-of-Fit dengan melihat nilai deviance dan membandingkan pula nilai AIC dari setiap model.

Hipotesis:

$$H_0: \lambda_{ij}^{XY} = 0$$
$$H_1: \lambda_{ij}^{XY} \neq 0$$

Loglinear Model	Residual Deviance	AIC
S,E,I	11444	11529
SE,I	3567.7	3654.7
SI,EI	7134	7223
SE,IE,SI	2.854	93.853
SEI	-5.009. e^-11	92.999

Apabila dilihat dari hasil perbandingan residual deviance dan AIC, akan dipilih nilai terkecil. Oleh karena itu, mod1, yaitu (SE,SI,IE) dipilih sebagai model yang cocok. (SEI tidak dipilih karena nilai residual $(-)$).

```
> mod2 <- glm(Freq ~ Ejected * Injury * Safety,
+             data = safety.df, family = poisson)
> summary(mod2)

Call:
glm(formula = Freq ~ Ejected * Injury * Safety, family = poisson,
    data = safety.df)

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

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      7.00760    0.03008 232.943  < 2e-16 ***
Ejectedno                        5.91902    0.03012 196.493  < 2e-16 ***
Injuryfatal                     -4.36854    0.26895 -16.243  < 2e-16 ***
Safetynone                       1.43141    0.03348  42.748  < 2e-16 ***
Ejectedno:Injuryfatal           -2.37806    0.27278  -8.718  < 2e-16 ***
Ejectedno:Safetynone            -2.39186    0.03362 -71.153  < 2e-16 ***
Injuryfatal:Safetynone           2.13812    0.27306   7.830 4.87e-15 ***
Ejectedno:Injuryfatal:Safetynone -0.44197    0.27863  -1.586    0.113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  1.6249e+06  on 7  degrees of freedom
Residual deviance: -5.0009e-11  on 0  degrees of freedom
AIC: 92.999

Number of Fisher Scoring iterations: 3
```

Saturated Model (bentuk acuan):

$$\log \mu_{ijk} = \lambda + \lambda_i^S + \lambda_j^E + \lambda_k^I + \lambda_{ij}^{SE} + \lambda_{ik}^{SI} + \lambda_{jk}^{EI} + \lambda_{ijk}^{SEI}$$

$$log\ \mu_{ijk} = 7.0076 + 1.43141S + 5.91902E - 4.36854I - 2.39186SE + 2.13812SI - 2.37806EI - 0.44197SEI$$

Model yang diajukan:

$$\log \mu_{ijk} = \lambda + \lambda_i^S + \lambda_j^E + \lambda_k^I + \lambda_{ij}^{SE} + \lambda_{ik}^{SI} + \lambda_{jk}^{EI}$$

$$\log \mu_{ijk} = 7.00137 + 1.43913S + 5.92527E - 3.936315I$$
$$- 2.39964SE + 1.71732SI - 2.79779EI$$

5. [**C4, 5 points**] Berikan interpretasi dari hasil model terbaik pada soal 3 di atas.

```
> mod1 <- glm(Freq ~ (Ejected + Injury + Safety)^2,
+             data = safety.df, family = poisson)
> summary(mod1)

Call:
glm(formula = Freq ~ (Ejected + Injury + Safety)^2, family = poisson,
    data = safety.df)

Deviance Residuals:
       1        2        3        4        5        6        7        8
 0.20704  -0.01071  -1.59987   0.31400  -0.10095   0.01731   0.30951  -0.21583

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           7.00137    0.02992  233.99   <2e-16 ***
Ejectedno             5.92527    0.02996  197.76   <2e-16 ***
Injuryfatal          -3.96315    0.06944  -57.07   <2e-16 ***
Safetynone            1.43913    0.03321   43.33   <2e-16 ***
Ejectedno:Injuryfatal -2.79779    0.05526  -50.63   <2e-16 ***
Ejectedno:Safetynone  -2.39964    0.03334  -71.97   <2e-16 ***
Injuryfatal:Safetynone 1.71732    0.05402   31.79   <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: 1.6249e+06  on 7  degrees of freedom
Residual deviance: 2.8540e+00  on 1  degrees of freedom
AIC: 93.853

Number of Fisher Scoring iterations: 3
```

Berdasarkan model summary pada mod1, diperoleh estimasi parameter dari model yang diajukan yaitu:

$$\log \mu_{ijk} = \lambda + \lambda_i^S + \lambda_j^E + \lambda_k^I + \lambda_{ij}^{SE} + \lambda_{ik}^{SI} + \lambda_{jk}^{EI}$$
$$\log \mu_{ijk} = 7.00137 + 1.43913S + 5.92527E - 3.936315I$$
$$- 2.39964SE + 1.71732SI - 2.79779EI$$

Dapat dilihat bahwa p-value pada masing-masing koefisien cukup signifikan, sehingga mengindikasikan adanya hubungan antar variabel, dimana hubungan tersebut dapat dilihat melalui odds ratio yang akan dihitung selanjutnya.

Untuk mengestimasi odds ratio, akan digunakan rumus $\hat{\theta} = exp\,\hat{\lambda}$

- Variabel Safety Equipment dan Whether Ejected - (SE)

$$\hat{\lambda} = -2.3996$$
$$\hat{\theta} = exp(-2.3996) = 0.091$$

Maka dapat disimpulkan bahwa peluang pengemudi tidak menggunakan seatbelt tidak akan terlontar keluar adalah 9.1% dibandingkan dengan peluang pengemudi yang menggunakan seatbelt.

- Variabel Safety Equipment dan Injury - (SI)

$$\hat{\lambda} = 1.7173$$
$$\hat{\theta} = exp(1.7173) = 5.57$$

Maka dapat disimpulkan bahwa peluang pengemudi tidak menggunakan seatbelt akan mengalami cedera fatal adalah 557% dibandingkan dengan peluang pengemudi yang menggunakan seatbelt.

- Variabel Whether Ejected dan Injury - (EI)

$$\hat{\lambda} = -2.7978$$
$$\hat{\theta} = exp(-2.7978) = 0.061$$

Maka dapat disimpulkan bahwa peluang pengemudi tidak terlontar keluar akan mengalami cedera fatal adalah 6.1% dibandingkan dengan peluang pengemudi yang terlontar keluar.

6. Jika ingin dilakukan analisis hubungan antara penggunaan seat belt dan whether ejected (yes/no) terhadap fatal atau tidaknya injury yang dialami penumpang suatu kendaraan saat terjadi kecelakaan; menggunakan model regresi logistik:

a. [**C4, 5 points**] Tuliskan model regresi yang sesuai untuk data tersebut (sesuaikan dengan rekomendasi model pada soal 4).
Model Loglinier yang mengandung SE ekuivalen dengan model logistic dengan I sebagai respons, serta S dan E sebagai variabel penjelas. Hal tersebut dapat terjadi disebabkan karena variable penjelas merupakan variabel kategorik sehingga model LogLinier dengan model logistic tersebut adalah ekuivalen.

```
Call:
glm(formula = Injury ~ Ejected + Safety, family = binomial, data = safety.df,
    weights = Freq)

Deviance Residuals:
     1      2      3      4      5      6      7      8
 -6.45  -30.85   10.56   80.82  -30.50  -44.98   48.29  100.90

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.96315    0.06944  -57.07   <2e-16 ***
Ejected0    -2.79779    0.05526  -50.63   <2e-16 ***
Safety0      1.71732    0.05401   31.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26669  on 7  degrees of freedom
Residual deviance: 23104  on 5  degrees of freedom
AIC: 23110

Number of Fisher Scoring iterations: 8
```

Dari soal nomor 4, didapat rekomendasi model loglinear berikut

$$\log \mu_{ijk} = 7.00137 + 1.43913S + 5.92527E - 3.936315I - 2.39964SE + 1.71732SI - 2.79779EI$$

Dengan hasil model tersebut didapatkan model regresi logistik yang sesuai adalah

$$\log\frac{\pi}{1-\pi} = -3.96315 - 2.79778E + 1.71732\,S$$

b. [**C4, 10 points**] Interpretasikan hasil yang diperoleh (lampirkan codes/screenshoot pengolahan data). Seberapa besar efek dari penggunaan seatbelt terhadap fatal/tidaknya injury yang dialami? Bagaimana dengan akibat dari “ejected”?

```
> exp(1.71732)
[1] 5.569582
```

Didapatkan nilai *odds ratio* bersyaratnya adalah 5.569586. Artinya, peluang penumpang yang tidak menggunakan *seatbelt* akan mengalami cedera fatal yaitu sebesar 5.57 kali lebih besar dibandingkan dengan peluang penumpang yang menggunakan *seatbelt.*

```
> exp(-2.79779)
[1] 0.0609446
```

Didapatkan nilai *odds ratio* bersyaratnya adalah 0.06094433. Artinya, peluang penumpang tidak terlontar keluar akan mengalami cedera fatal yaitu sebesar 0.061 kali lebih besar dibandingkan dengan peluang penumpang yang terlontar keluar.

c. [**C4, 5 points**] Apakah efek dari penggunaan seatbelt terhadap fatal/tidaknya injury bergantung pada terjadi “ejected” atau tidak? Jelaskan.
Ya. Efek dari penggunaan *seatbelt* terhadap fatal atau tidaknya *injury* memiliki ketergantungan pada terjadinya “*ejected*” atau tidak. Dengan kata lain, peluang pengemudi terlontar atau tidak akan berpengaruh pada efek dari penggunaan seatbelt terhadap terjadinya fatal *injury* atau tidak. Hal ini sesuai dengan informasi yang diperoleh dari hasil interpretasi pada bagian b, yakni:

- Penumpang yang tidak menggunakan *seatbelt* akan mengalami cedera fatal yaitu sebesar 5.57 kali lebih besar dibandingkan dengan peluang penumpang yang menggunakan *seatbelt.*

- Penumpang yang tidak menggunakan *seatbelt* memiliki risiko terlontar 11 kali lipat lebih besar dibandingkan dengan penumpang yang menggunakan *seatbelt*. (Dilihat dari hasil loglinear karena pada hasil mode logistic tidak keluar hubungan safety dengan ejected)

```
> exp(2.39964)
[1] 11.01921
```

d. [**C4, 10 points**] Bandingkan hasil (informasi) yang diperoleh pada bagian b) dengan hasil pada soal 4). Apakah sesuai? Apa kekurangan/kelebihan dari masing-masing metode analisis tersebut?

Hasil dari nilai odds ratio SI (Safety-Injury) dan EI (Ejected-Injury) keduanya sama. Namun, pada hasil dari soal b, hanya didapatkan odds ratio dari variabel Safety terhadap Injury dan Ejected terhadap Injury. Akan tetapi, tidak terdapat odds ratio antara Safety dengan Ejected. Hal ini terjadi karena variabel Injury pada model logistic menjadi variabel respons.

```
Call:
glm(formula = Injury ~ (Ejected + Safety)^2, family = binomial,
    data = safety.df, weights = Freq)

Deviance Residuals:
      1       2       3       4       5       6       7       8
 -5.275  -31.071   11.076   80.736  -30.727  -44.828   48.151  100.968

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.3685     0.2689 -16.243  < 2e-16 ***
Ejected0          -2.3781     0.2728  -8.718  < 2e-16 ***
Safety0            2.1381     0.2731   7.830 4.87e-15 ***
Ejected0:Safety0  -0.4420     0.2786  -1.586    0.113
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26669  on 7  degrees of freedom
Residual deviance: 23101  on 4  degrees of freedom
AIC: 23109

Number of Fisher Scoring iterations: 8
```

Apabila dihitung pun, koefsien yang didapatkan berbeda dan tidak signifikan.

Kelebihan model LogLinear adalah memang dikhususkan untuk variabel-variabel yang dimana nilai X dan Y nya adalah kategorik dan juga kita dapat mengetahui hubungan / asosiasi yang dapat dicari dengan *odds ratio* dari ketiga variabel. Hal ini tidak bisa kita dapatkan apabila kita menggunakan model logistic.

Bagian 3. Penutup [**C4, 5 points**]
[Tuliskan kesimpulan apa yang kalian dapatkan dari penugasan ini, kaitkan dengan permasalahan yang diajukan pada **Bagian 1**.**]**

Dari data yang diberikan, dapat diketahui bahwa terdapat asosiasi antara penggunaan seatbelt terhadap injury. Diketahui juga bahwa terdapat perbedaan tingkat resiko dari menggunakan seatbelt dan tidak terhadap injury dalam kondisi ejected. Pada data ini, digunakan model loglin seperti saran pada bagian 1 dan didapatkan perbandingan model terbaik yang akan digunakan. Kemudian, didapatkan pula model loglin yang setara dengan mode logistik dimana I (Injury) pada data berupa respon dan S (Safety) serta E (Ejected) berupa variabel penjelas.

Selain itu, didapatkan pula perhitungan antara asosiasi dari ketiga variabel kategorik, dan bagaimana pengaruh variabel tersebut.

Bagian 4. Lampiran

Tuliskan codes R/Python, atau screenshot proses pengolahan data.

```
safety <- array(data = c(1105, 411111, 14, 483, 4624, 157342, 497,
1008),
                dim = c(2,2,2),
                dimnames = list("Ejected" = c("yes","no"),
                                "Injury" = c("nonfatal","fatal"),
                                "Safety" = c("seatbelt","none")))
safety

ftable(safety, row.vars = c("Safety","Ejected"))

addmargins(safety)

prop.table(safety, margin = c(1,3))

safety.df <- as.data.frame(as.table(safety))
safety.df[,-4] <- lapply(safety.df[,-4], relevel, ref = "no")
safety.df

mod0 <- glm(Freq ~ Ejected + Injury + Safety,
            data = safety.df, family = poisson)
summary(mod0)

pchisq(deviance(mod0), df = df.residual(mod0), lower.tail = F)

cbind(mod0$data, fitted(mod0))

exp(coef(mod0)[3])

margin.table(safety, margin = 2)/sum(margin.table(safety, margin =
2))

mod1 <- glm(Freq ~ (Ejected + Injury + Safety)^2,
            data = safety.df, family = poisson)
summary(mod1)

pchisq(deviance(mod1), df = df.residual(mod1), lower.tail = F)

cbind(mod1$data, fitted(mod1))

exp(coef(mod1)["Injuryfatal:Safetynone"])

exp(confint(mod1, parm = c("Ejectedno:Injuryfatal",
                           "Ejectedno:Safetynone",
                           "Injuryfatal:Safetynone")))

mod2 <- glm(Freq ~ Ejected * Safety * Injury,
            data = safety.df, family = poisson)
summary(mod2)
deviance(mod2)
cbind(mod2$data, fitted(mod2))

anova(mod1, mod2)

mod0 <- glm(Freq ~ Ejected + Injury + Safety,
            data = safety.df, family = poisson)
```

```
summary(mod0)
cbind(mod0$data, fitted(mod0))

mod1 <- glm(Freq ~ (Ejected + Injury + Safety)^2,
            data = safety.df, family = poisson)
summary(mod1)
cbind(mod1$data, fitted(mod1))

mod2 <- glm(Freq ~ Ejected * Safety * Injury,
            data = safety.df, family = poisson)
summary(mod2)
cbind(mod2$data, fitted(mod2))

mod3 <- glm(Freq ~ (Safety*Injury)+(Ejected*Injury),
            data = safety.df, family = poisson)
summary(mod3)
cbind(mod3$data, fitted(mod3))

mod4 <- glm(Freq ~ (Safety*Ejected)+Injury,
            data = safety.df, family = poisson)
summary(mod4)
cbind(mod4$data, fitted(mod4))

model<-(fit.glm<-glm(Freq~.^2, data=safety.df, family=poisson))
summary(model)


model<-(fit.glm<-glm(Freq~.^2, data=safety.df, family=poisson))
summary(model)

##logistic model
safety <- array(data = c(1105, 411111, 14, 483, 4624, 157342, 497,
1008),
                dim = c(2,2,2),
                dimnames = list("Ejected" = c(1,0),
                                "Injury" = c(0,1),
                                "Safety" = c(1,0)))
safety.df <- as.data.frame(as.table(safety))
model<-glm(Injury~Ejected+Safety, data=safety.df,
family=binomial,weights= Freq)
summary(model)
```