P1

Problem 1: (a) Number of accident happen in the west = P(accident happens in the west) * total number of work force = P(the work force is in the west factory) * P(there is an accident) * total number of work force = 1306/1979 * 59/1979 *1979 = 38.9. So the number of accident happened in the west factory is around 39.

```
dat = expand.grid(factory=c("East", "West"), accident=c("No", "Yes"))
dat$y = c(645,1275, 28,31)
tab = matrix(dat$y, nrow=2,
  dimnames=list(factory=c("East", "West"), accident=c("No", "Yes")))

#(b)
chisq.test(tab)$expected
```

```
##         accident
## factory        No      Yes
##    East  652.9358 20.06417
##    West 1267.0642 38.93583
```

(c) Let 1 represent the west factory or accident and 2 represent the east factory or no accident. m11 = pi_1+ * pi_+1 * 1979 m12 = pi_1+ * pi_+2 * 1979 m21 = pi_2+ * pi_+1 * 1979 m22 = pi_2+ * pi_+2 * 1979 To generalize, m_ij = pi_i+ * pi_+j * n.

(d) log(m11) = log(pi_1+)+ log(pi_+1)+log(1979) log(m12) = log(pi_1+)+ log(pi_+2) + log(1979) log(m21) = log(pi_2+)+ log(pi_+1) +log(1979) log(m22) = log(pi_2+)+ log(pi_+2) +log(1979) In terms of generalized form, log(m_ij) = log(pi_i+) + log(pi_+j) + log(n)

(e) log(pi_ij) = log(pi_+i) + log(pi_+j)

HW07q2

Samuel Swain

2022-11-30

Question 2

Data

```
dat = expand.grid(factory=c("East", "West"), accident=c("No", "Yes"))
dat$y = c(645,1275, 28,31)
tab = matrix(dat$y, nrow=2,
dimnames=list(factory=c("East", "West"), accident=c("No", "Yes")))
```

2(a)

```
fit_2a = glm(y ~ factory + accident, poisson, dat)
fit_2a
```

```
##
## Call:  glm(formula = y ~ factory + accident, family = poisson, data = dat)
##
## Coefficients:
## (Intercept)  factoryWest  accidentYes
##       6.481        0.663       -3.483
##
## Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
## Null Deviance:       2423
## Residual Deviance: 4.678     AIC: 38.43
```

2(b)

```
predict(fit_2a, newdata = data.frame(factory = factor("West"),
   accident = factor("Yes")), type = "response")
```

```
##        1
## 38.93583
```

To attain the result manually, we can use the equation bellow:

$$e^{(6.481+0.663+-3.483)} = 38.9$$

2(c)

```
fit_2c = glm(y ~ factory*accident, poisson, dat)
fit_2c
```

```
##
## Call:  glm(formula = y ~ factory * accident, family = poisson, data = dat)
##
## Coefficients:
##             (Intercept)              factoryWest              accidentYes
##                  6.4693                   0.6815                  -3.1370
## factoryWest:accidentYes
##                 -0.5797
##
## Degrees of Freedom: 3 Total (i.e. Null);  0 Residual
## Null Deviance:       2423
## Residual Deviance: 1.061e-13     AIC: 35.75
```

```
predict(fit_2c, newdata = data.frame(factory = factor("West"),
   accident = factor("Yes")), type = "response")
```

```
##  1
## 31
```

To attain the result manually, we can use the equation bellow:

$$e^{(6.4693+0.6815-3.1370-0.5797)} = 31.0$$

2(d)

We get a residual deviance of 0 because our predicted model is the saturated model. Our model can adjust to fit any row in the data frame perfectly. Thus, when we subtract the log-likelihood of the saturated model from the log-likelihood of the predicted model, we will get zero.

2(e)

```
summary(fit_2c)
```

```
##
## Call:
## glm(formula = y ~ factory * accident, family = poisson, data = dat)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              6.46925    0.03937 164.299   <2e-16 ***
## factoryWest              0.68145    0.04832  14.103   <2e-16 ***
## accidentYes             -3.13705    0.19304 -16.251   <2e-16 ***
## factoryWest:accidentYes -0.57967    0.26515  -2.186   0.0288 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.4235e+03  on 3  degrees of freedom
## Residual deviance: 1.0614e-13  on 0  degrees of freedom
## AIC: 35.749
##
## Number of Fisher Scoring iterations: 3
```

The z-value for the interaction term is -2.186. This is less than -1.645. We can reject the null that $\beta_{factory*accident} = 0$ and conclude $\beta_{factory*accident} \neq 0$.

2(f)

The result from question 2 part e tells us the west factory on average is less likely to have accidents.

2(g)

```
d = drop1(fit_2c, test="Chisq")
d
```

```
## Single term deletions
##
## Model:
## y ~ factory * accident
##                  Df Deviance    AIC   LRT Pr(>Chi)
## <none>                0.000 35.749
## factory:accident  1    4.678 38.427 4.678  0.03055 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
p_val_2g = d$`Pr(>Chi)`[2]
chi_sq_2g = qchisq(p_val_2g, 1, lower.tail = FALSE)
```

Using the likelihood ratio test to evaluate the interaction, we get a chi-squared value of 4.6779804 and a p-value of 0.0305516.

2(h)

```
east_accident_0 = predict(fit_2c, newdata = data.frame(factory = factor("East"), accident = factor("No")), t
ype = "link")
east_accident_1 = predict(fit_2c, newdata = data.frame(factory = factor("East"), accident = factor("Yes")),
 type = "link")

west_accident_0 = predict(fit_2c, newdata = data.frame(factory = factor("West"), accident = factor("No")), t
ype = "link")
west_accident_1 = predict(fit_2c, newdata = data.frame(factory = factor("West"), accident = factor("Yes")),
 type = "link")

log_odds_east_2h = east_accident_1 - east_accident_0
log_odds_west_2h = west_accident_1 - west_accident_0
```

- The log odds of an accident in the east is -3.1370459
- The log odds of an accident in the west is -3.7167143

2(i)

```
c = log_odds_east_2h
d = log_odds_west_2h - log_odds_east_2h
cat("c: ", c, "\n", "d: ", d)
```

```
## c:  -3.137046
##  d:  -0.5796684
```

$$log(\frac{\pi_{1|i}}{1-\pi_{1|i}}) = log(\frac{\pi_{1|i}}{\pi_{0|i}})$$
$$= log(m_{i0}) + log(\frac{m_{i1}}{m_{i0}}) * west$$
$$= -3.1370 \ - \ 0.5797 * west$$

2(j)

```
fit_2j = glm(accident ~ factory, binomial, dat, weights = y)
summary(fit_2j)
```

```
##
## Call:
## glm(formula = accident ~ factory, family = binomial, data = dat,
##     weights = y)
##
## Deviance Residuals:
##       1       2       3       4
## -7.404  -7.827  13.344  15.229
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.1370     0.1930 -16.251   <2e-16 ***
## factoryWest  -0.5797     0.2651  -2.186   0.0288 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 530.73  on 3  degrees of freedom
## Residual deviance: 526.06  on 2  degrees of freedom
## AIC: 530.06
##
## Number of Fisher Scoring iterations: 6
```

2(k)

I would expect the estimated logistic regression to be just the intercept: $log(\frac{\pi_{1|i}}{1-\pi_{1|i}}) = -3.1370$

Problem #3

(a) $\log \pi_{ik} = \alpha_k^T x_i - \log z$

$\log z = \alpha_k^T x_i - \log \pi_{ik}$

$\exp(\log z) = \exp(\alpha_k^T x_i - \log \pi_{ik})$

$z = \exp(\alpha_k^T x_i) / \pi_{ik}$ for $k = 1, 2, \cdots, K$

$\because \sum_{k=1}^{K} \pi_{ik} = 1$

$\therefore z = \sum_{k=1}^{K} \exp(\alpha_k^T x_i)$

(b) $\log \pi_{ik} = \alpha_k^T x_i - \log z$

$\exp(\log \pi_{ik}) = \exp(\alpha_k^T x_i - \log z)$

$\pi_{ik} = \exp(\alpha_k^T x_i) / z$

(c) $\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \beta_k^T x_i$

$\log(\pi_{ik}) - \log(\pi_{i1}) = \beta_k^T x_i$, for $k = 2, \cdots, K$

$\log \pi_{ik} = \beta_k^T x_i + \log \pi_{i1}$, for $k = 2, \cdots, K$

$\pi_{ik} = \exp(\beta_k^T x_i) \cdot \pi_{i1}$, for $k = 2, \cdots, K$

$\because \pi_{ik} = \exp(\alpha_k^T x_i) / z$ from 3b

$\therefore \exp(\beta_k^T x_i) \cdot \pi_{i1} = \exp(\alpha_k^T x_i) / z$

$\exp(\beta_k^T x_i) = \exp(\alpha_k^T x_i) / z \cdot \pi_{i1}$

$\beta_k^T x_i = \alpha_k^T x_i - \log(z \cdot \pi_{i1})$

$\beta_k^T = \alpha_k^T - \log(z \cdot \pi_{i1})$

$\beta_k^T = \alpha_k^T - (\log(z) + \log \pi_{i1})$

Also from previous: $\log \gamma_{ik} = \alpha_k^T x_i - \log Z$

$\therefore$ for $k = 1$: $\log \gamma_{i1} = \alpha_1 x_i - \log Z$

$\log \gamma_{i1} + \log Z = \alpha_1 x_i$

$\therefore \beta_k^T = \alpha_k^T - \alpha_1 x_i$

$\therefore \beta_k = \alpha_k - \alpha_1 x_i$, where $\alpha_1 x_i$ is a constant

MSiA-401-hw7-q4

2022-11-29

```
setwd("~/Desktop/MSiA-401-hw7")
desert = read.csv("desert.csv",header = T)
# omit NA
desert2 = na.omit(desert)
head(desert2,10)
```

```
##    FIPS newsPub  age    pop BAhigher income raceBlack race_Hisp digDistress
## 1  1003       4 42.6 203360     30.7  52562       9.5       4.4    28.58287
## 2  1005       1 39.7  26201     12.0  33368      47.8       4.2    50.65492
## 3  1007       1 39.8  22580     13.2  43404      22.0       2.4    52.60406
## 4  1009       1 40.9  57667     13.1  47412       1.5       9.0    39.09295
## 5  1011       1 40.8  10478     13.4  29655      75.6       0.3    62.23687
## 6  1013       1 40.7  20126     16.1  36326      44.7       0.3    48.72265
## 7  1015       1 39.1 115527     17.9  43686      20.4       3.6    32.97889
## 8  1017       2 43.0  33895     13.3  37342      39.3       2.2    46.54169
## 9  1019       1 46.1  25855     12.5  40041       5.0       1.6    43.37013
## 10 1021       2 38.9  43805     15.1  43501       9.5       7.7    51.71853
```

```
fit1 = glm(newsPub ~ . -FIPS, poisson, data = desert2)
summary(fit1)
```

```
##
## Call:
## glm(formula = newsPub ~ . - FIPS, family = poisson, data = desert2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.9405  -0.6614  -0.2605   0.3281   9.8664
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.058e-01  1.570e-01   3.859 0.000114 ***
## age         -1.876e-03  2.654e-03  -0.707 0.479574
## pop          3.890e-07  9.183e-09  42.360  < 2e-16 ***
## BAhigher     1.416e-02  1.944e-03   7.283 3.27e-13 ***
## income       3.734e-06  1.360e-06   2.745 0.006049 **
## raceBlack    4.300e-03  1.102e-03   3.902 9.54e-05 ***
## race_Hisp    3.278e-03  1.069e-03   3.068 0.002159 **
## digDistress -1.627e-02  1.977e-03  -8.229  < 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: 6123.8  on 3141  degrees of freedom
## Residual deviance: 3693.6  on 3134  degrees of freedom
## AIC: 10792
##
## Number of Fisher Scoring iterations: 5
```

First we fit a Poisson model fit1 by simply use all variables except FIPS, we observe the deviance for the model is 3693.6 and AIC is 10792.

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

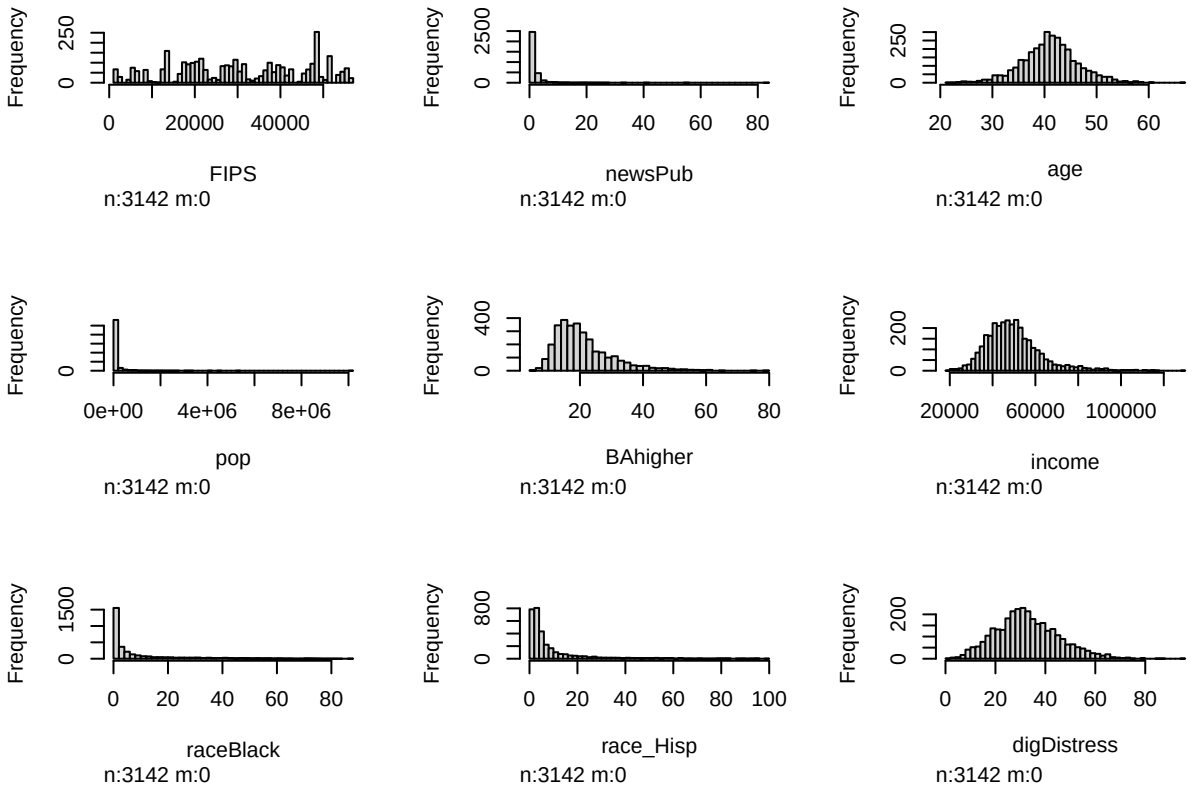
```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
hist.data.frame(desert2)
```



By observing the data distribution for all variables, we find variables pop, raceBlack, race_Hisp have very non-symmetric distribution. In that case, we will transform these three varibles by taking log in the new model.

Besides, since digDistress is considered as a pip (determined by pop and income), we will not include digDistress in the new model.

```
fit2 = glm(newsPub ~ age + log(pop) + BAhigher + income + log(raceBlack+1) + log(race_Hisp+1), poisson,
summary(fit2)
```

```
##
## Call:
## glm(formula = newsPub ~ age + log(pop) + BAhigher + income +
##     log(raceBlack + 1) + log(race_Hisp + 1), family = poisson,
##     data = desert2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4815  -0.6708  -0.1023   0.3972  12.5225
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.133e+00  1.746e-01 -29.395  < 2e-16 ***
## age                 2.099e-02  2.810e-03   7.469 8.11e-14 ***
## log(pop)            4.548e-01  1.134e-02  40.106  < 2e-16 ***
## BAhigher            2.727e-03  1.940e-03   1.405   0.1599
## income              2.613e-06  1.290e-06   2.025   0.0428 *
```

```
## log(raceBlack + 1) -1.109e-01  1.320e-02  -8.407  < 2e-16 ***
## log(race_Hisp + 1)  3.060e-02  1.588e-02   1.927   0.0540 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6123.8  on 3141  degrees of freedom
## Residual deviance: 2957.6  on 3135  degrees of freedom
## AIC: 10054
##
## Number of Fisher Scoring iterations: 5
```

We observe the new model fit2 improves significantly because deviance and AIC of fit2 is 2957.6 and 10054, which is less than 3693.6 and 10792.

We will then use drop1 to see if dropping any other variables could improve the model.

```
drop1(fit2)
```

```
## Single term deletions
##
## Model:
## newsPub ~ age + log(pop) + BAhigher + income + log(raceBlack +
##     1) + log(race_Hisp + 1)
##                     Df Deviance   AIC
## <none>                   2957.6 10054
## age                  1   3012.9 10107
## log(pop)             1   4693.6 11788
## BAhigher             1   2959.6 10054
## income               1   2961.7 10056
## log(raceBlack + 1)   1   3029.8 10124
## log(race_Hisp + 1)   1   2961.3 10055
```

```
summary(fit2)
```

```
##
## Call:
## glm(formula = newsPub ~ age + log(pop) + BAhigher + income +
##     log(raceBlack + 1) + log(race_Hisp + 1), family = poisson,
##     data = desert2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4815  -0.6708  -0.1023   0.3972  12.5225
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.133e+00  1.746e-01 -29.395  < 2e-16 ***
## age                 2.099e-02  2.810e-03   7.469 8.11e-14 ***
## log(pop)            4.548e-01  1.134e-02  40.106  < 2e-16 ***
## BAhigher            2.727e-03  1.940e-03   1.405   0.1599
## income              2.613e-06  1.290e-06   2.025   0.0428 *
```

```
## log(raceBlack + 1) -1.109e-01  1.320e-02  -8.407  < 2e-16 ***
## log(race_Hisp + 1)  3.060e-02  1.588e-02   1.927   0.0540 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6123.8  on 3141  degrees of freedom
## Residual deviance: 2957.6  on 3135  degrees of freedom
## AIC: 10054
##
## Number of Fisher Scoring iterations: 5
```

We observe dropping any of variables would not improve the model because dropping any variables would increase deviance of the model. But we notice variables BAhigher and log(race_Hisp + 1) is non-significant (have p-value greater than 0.05), dropping them also barely affect deviance and AIC. In that case, we drop variables BAhigher and log(race_Hisp + 1).

```
fit3 = glm(newsPub ~ age + log(pop) + income + log(raceBlack+1), poisson, data = desert2)
summary(fit3)
```

```
##
## Call:
## glm(formula = newsPub ~ age + log(pop) + income + log(raceBlack +
##     1), family = poisson, data = desert2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4642  -0.6701  -0.0997   0.4025  12.5988
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.136e+00  1.703e-01 -30.155  < 2e-16 ***
## age                 1.942e-02  2.719e-03   7.143 9.11e-13 ***
## log(pop)            4.661e-01  1.020e-02  45.689  < 2e-16 ***
## income              3.915e-06  9.378e-07   4.175 2.98e-05 ***
## log(raceBlack + 1) -1.125e-01  1.310e-02  -8.590  < 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: 6123.8  on 3141  degrees of freedom
## Residual deviance: 2962.8  on 3137  degrees of freedom
## AIC: 10055
##
## Number of Fisher Scoring iterations: 5
```

From the model above, we can see predictor variables age, log(pop), income have positive effect on dependent variable newsPub, whereas predictor variable log(raceBlack + 1) has negative effect on newsPub. This makes common sense since older people tend to read more newspaper; people read newspaper will increase as the population increase; people with higher income can afford buying newspaper than people with less income.