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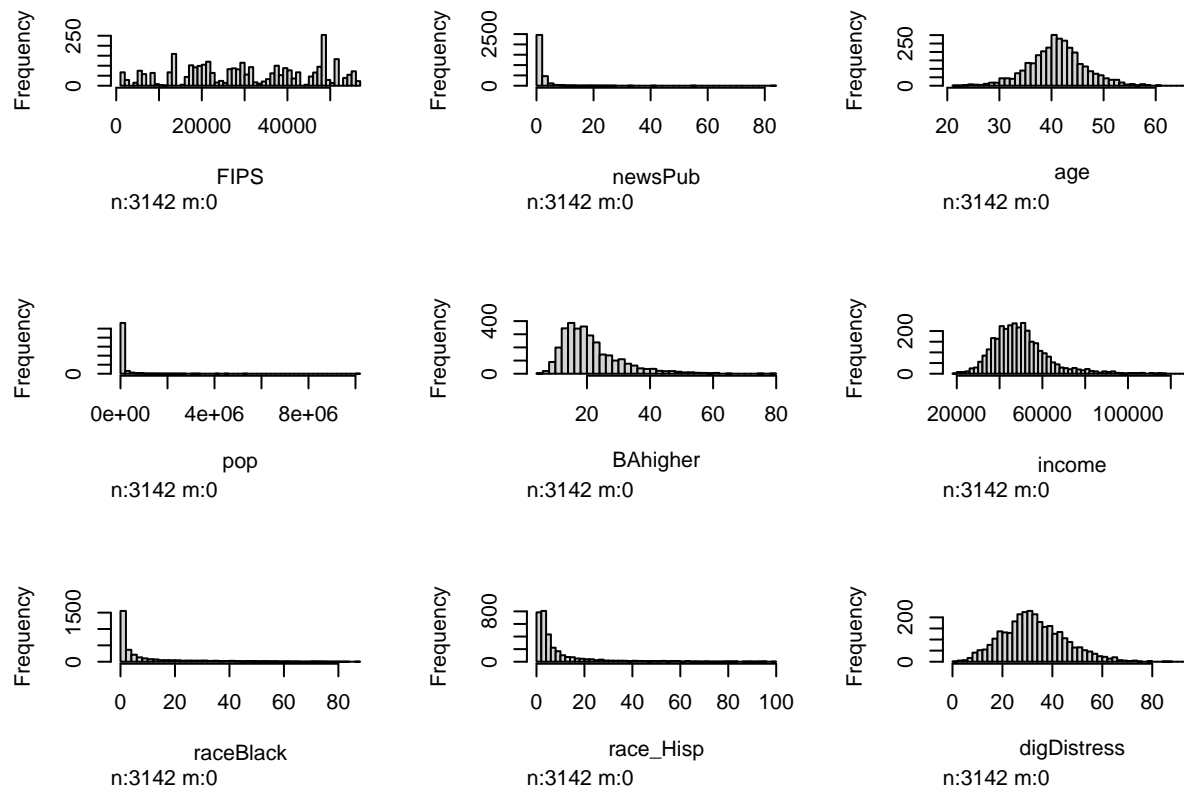
```
## Loading required package: ggplot2
```

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

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

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

Frequency | 0 250

0   20000   40000

FIPS

n:3142 m:0

Frequency | 0 2500

0   20   40   60   80

newsPub

n:3142 m:0

Frequency | 0 250

20 30 40 50 60

age

n:3142 m:0

Frequency | 0

0e+00   4e+06   8e+06

pop

n:3142 m:0

Frequency | 0 400

20   40   60   80

BAhigher

n:3142 m:0

Frequency | 0 200

20000   60000   100000

income

n:3142 m:0

Frequency | 0 1500

0   20   40   60   80

raceBlack

n:3142 m:0

Frequency | 0 800

0   20   40   60   80   100

race_Hisp

n:3142 m:0

Frequency | 0 200

0   20   40   60   80

digDistress

n:3142 m:0

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 *
```

3

```
## 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.