566 Project Model Fitting, Selection, and Diagnostics

Kaelan Yu

May 18, 2021

0. Preliminaries

```
# setwd("C:/Users/Kaelan/Desktop/stat-566-causal-inference/code")

# access to stepAIC() function
library(MASS)

# make dummy variables
library(fastDummies)

# access variance inflation factor function
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
# make correlation plot
library(corrplot)
```

```
## corrplot 0.88 loaded
```

```
source('566_exploratoryDataAnalysis.R')
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:ggpubr':
##
##     mutate
```

```
# convert variables to factors
for (i in 1:k) {
  dat[, i] <- as.factor(dat[, i])
}
```

1. Fitting our Model (Logistic Regression)

```
# full model
glm.full <- glm(h1n1_vaccine ~ ., family = binomial(link = 'logit'), data = dat)
```

2. Variable Selection using Bayesian Information Criterion (BIC)

penalty $\propto \log(n)$

```
best.BIC <- stepAIC(glm.full, direction = "both", trace = FALSE, k = log(n))

best.BIC.variables <- attr(best.BIC$terms, "term.labels")

dat.best <- dat[c("h1n1_vaccine", best.BIC.variables)]
```

(AIC: penalty $\propto 2$)

3. Model Diagnostics - Analyzing Multicollinearity Issues

```
# make all categorical variables binary dummy variables
# k - 1 dummy variables for a variable of level k to avoid multicollinearity issues
dat.best.dummy <- dummy_cols(dat.best, remove_first_dummy = TRUE, remove_selected_columns = TRUE)


# variance inflation factors (VIF)
vif(best.BIC) # VIF all below 5 so no notable multicollinearity issues
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## h1n1_knowledge             1.097742  2        1.023588
## behavioral_large_gatherings 1.079967  1        1.039215
## doctor_recc_h1n1           1.773874  1        1.331868
## doctor_recc_seasonal       1.832006  1        1.353516
## child_under_6_months       1.018385  1        1.009151
## health_worker              1.062424  1        1.030740
## health_insurance           1.065621  1        1.032289
## opinion_h1n1_vacc_effective 1.195438  4        1.022565
## opinion_h1n1_risk          1.449768  4        1.047520
## opinion_h1n1_sick_from_vacc 1.294465  4        1.032788
## sex                        1.070007  1        1.034412
## marital_status             1.033180  1        1.016455
## seasonal_vaccine           1.145248  1        1.070163
```

34 → 13 (p)
26707 → 11794 (n)

```
# correlation plots
corrplot(cor(dat.best.dummy))
```