INDE546_HW5

Michael Shieh

2/6/2020

Exercise 1

First, import dataset as survey:

```
survey <- read.csv("Class_Survey_W20.csv", header = TRUE)
```

The dependent variable I have chosen for this logistic regression model is **whether the participants have involved in a car crash or not in the past 5 years**, where there are two levels (Yes and No).

```
survey <- survey %>%
  rename(CarCrash = In.the.past.5.years..how.many.times.have.you.bee
n.involved.in.a.car.crash..) %>%
  drop_na(CarCrash) %>%
  mutate(CarCrash = ifelse(CarCrash == 0, 0, 1))
```

The first independent variable I have chosen is **how often do the participants go on campus every week**, where there are two levels (high and low). I chose this variable because I think if you stay at home more often, it is more unlikely that you will encounter a car crash.

```
survey <- survey %>% rename(OnCampusFreq = In.an.average.week..how.m
any.days.are.you.on.the.UW.campus.)
survey <- survey %>% drop_na(OnCampusFreq)
survey <- survey %>% mutate(OnCampusFreq = ifelse(grepl('Every', sur
vey$OnCampusFreq), 1, 0))

survey$OnCampusFreq <- factor(survey$OnCampusFreq, levels = c("0", "
1"), labels = c("Low", "High"))
```

The second independent variable I have chosen is **do participants sort wastes properly**, where there are two levels (Yes and No). I chose this variable because I think if people follow traffic laws, they should encounter less car crashes. If willing to follow laws, they might be more willing to sort wastes properly. I am curios to see the relationships between the two variables.

```
survey <- survey %>%
  rename(ProperlySort = How.often.....do.you.properly.sort.waste.int
o.trash..recyclables.and.compost.) %>%
  drop_na(ProperlySort) %>%
  mutate(ProperlySort = ifelse(ProperlySort == 'Never', 0, 1))

survey$ProperlySort <- factor(survey$ProperlySort, levels = c("0", "
1"), labels = c("No", "Yes"))
```

The third independent variable I have chosen is **the age of the participants**, where this is a continuous variable. I chose this variable because I believe age is definitely sharing a meaningful relationship with the dependent variable.

```
survey <- survey %>%
  rename(Age = How.old.are.you.) %>%
  filter(Age > 0)
```

The forth independent variable I have chosen is **how long do the participants own driver's licenses**, where this is a continuous variable. I chose this variable because I think that the longer you drive, you more experiences you have. Therefore, it should be more unlikely to encounter a car crash if you hold a driver's license longer.

```
survey <- survey %>%
  rename(DLAge = How.old.were.you.when.you.got.your.driver.s.license
.) %>%
  filter(DLAge > 0) %>%
  mutate(DL_Length = Age - DLAge)
```

The fifth independent variable I have chosen is **how often do participants drive weekly compare to other participants**, where there are two levels (above average and below average) I chose this variable because I think if you drive more frequently than others, it is more likely that you got involved in a car crash.

```
survey <- survey %>%
  rename(DriveFreq = On.average..how.many.days.in.a.week..out.of.7.d
ays..do.you.drive.) %>%
  filter(DriveFreq >= 0) %>%
  mutate(DriveFreq = ifelse(DriveFreq >= mean(DriveFreq), 1, 0))

survey$DriveFreq <- factor(survey$DriveFreq, levels = c("0", "1"), l
abels = c("BelowAvg", "AboveAvg"))
```

The sixth variable I have chosen is **the gender of the participants**, where there are two levels (male and female).

I chose this variable because I want to see if gender effect the frequency of involving in a car crash.

```
survey <- survey %>%
  rename(Gender = Are.you..1) %>%
  filter(Gender %in% c('Male', 'Female')) %>%
  mutate(Gender = case_when(Gender == 'Male' ~ 1, Gender == 'Female'
~ 0))

survey$Gender <- factor(survey$Gender, levels = c("0", "1"), labels
= c("Female", "Male"))
```

Next, I put all variables into the regression model.

```
LB_HW5 <- glm(CarCrash ~ Age + OnCampusFreq + ProperlySort + DL_Leng
th + DriveFreq + Gender, data = survey, family = binomial())
summary(LB_HW5)
```

```
##
## Call:
## glm(formula = CarCrash ~ Age + OnCampusFreq + ProperlySort +
##     DL_Length + DriveFreq + Gender, family = binomial(), data = s
urvey)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2169  -0.8458   0.4495   0.5685   1.8551
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          3.260415   1.412524   2.308   0.0210 *
## Age                  0.006479   0.046823   0.138   0.8899
## OnCampusFreqHigh     0.316322   0.307272   1.029   0.3033
## ProperlySortYes     -1.098622   1.058481  -1.038   0.2993
## DL_Length           -0.077845   0.046087  -1.689   0.0912 .
## DriveFreqAboveAvg   -2.488619   0.299298  -8.315   <2e-16 ***
## GenderMale          -0.307753   0.298402  -1.031   0.3024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.71  on 306  degrees of freedom
## Residual deviance: 288.69  on 300  degrees of freedom
## AIC: 302.69
##
## Number of Fisher Scoring iterations: 4
```

Exercise 2

Since age is a very insignificant variable, I remove this variable to see if I can produce a better model.

```
LB_HW5 <- glm(CarCrash ~ OnCampusFreq + ProperlySort + DL_Length + D
riveFreq + Gender, data = survey, family = binomial())
summary(LB_HW5)
```

```
##
## Call:
## glm(formula = CarCrash ~ OnCampusFreq + ProperlySort + DL_Length
+
##     DriveFreq + Gender, family = binomial(), data = survey)
##
## Deviance Residuals:
##     Min        1Q    Median        3Q       Max
## -2.2263  -0.8462   0.4485   0.5652   1.8517
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.38087    1.11324   3.037 0.002390 **
## OnCampusFreqHigh    0.31476    0.30692   1.026 0.305105
## ProperlySortYes    -1.08871    1.05552  -1.031 0.302329
## DL_Length          -0.07212    0.02023  -3.565 0.000364 ***
## DriveFreqAboveAvg -2.48813    0.29918  -8.317  < 2e-16 ***
## GenderMale         -0.31014    0.29794  -1.041 0.297894
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.71  on 306  degrees of freedom
## Residual deviance: 288.71  on 301  degrees of freedom
## AIC: 300.71
##
## Number of Fisher Scoring iterations: 4
```

After removing the variable age, I noticed that there is a new significant variable occured. Next, I removed all insignificant variable to see if I can furthur improve the model.

```
LB_HW5 <- glm(CarCrash ~ DL_Length + DriveFreq, data = survey, famil
y = binomial())
summary(LB_HW5)
```

```
##
## Call:
## glm(formula = CarCrash ~ DL_Length + DriveFreq, family = binomial
(),
##     data = survey)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2033  -0.8885   0.4600   0.5435   1.9414
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.33474    0.27283   8.558  < 2e-16 ***
## DL_Length         -0.07100    0.02029  -3.499 0.000466 ***
## DriveFreqAboveAvg -2.49262    0.29529  -8.441  < 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: 395.71  on 306  degrees of freedom
## Residual deviance: 291.44  on 304  degrees of freedom
## AIC: 297.44
##
## Number of Fisher Scoring iterations: 4
```

```
x <- data.frame("Variable Names" = c('Length of having drivers licen
se', 'Above average frequency of driving'), "coefficients" = LB_HW5$
coefficients[2:3], "z-value" = c('-3.499', '-8.441'), "Significance"
= c('Yes', 'Yes'))
x
```

```
##                                    Variable.Names coefficients
z.value
## DL_Length           Length of having drivers license  -0.07099598
-3.499
## DriveFreqAboveAvg Above average frequency of driving  -2.49262368
-8.441
##                   Significance
## DL_Length                  Yes
## DriveFreqAboveAvg          Yes
```

In this model, the length of holding a driver's license and the weekly driving frequency are the two significant variables effecting the dependent variable previously involving in car crashes or not. With the removal of the variables, the AIC decreases along the way. Including of all four remaining independent variables decreases the deviance by approximately 26.8% while sacrificing 4 degrees of freedom. Also, it takes 4 iterations to acheive a maximum likelihood estimate.

Next, I look at the relative risk and odds ratio.

```
table <- table(survey$DriveFreq, survey$CarCrash)
colnames(table) = c("No Crash", "Crash")
table
```

```
##
##             No Crash Crash
##   BelowAvg        26   160
##   AboveAvg        80    41
```

P(BelowAvg+Crash) = 160/186 = 0.8602, P(AboveAvg+Crash) = 41/121 = 0.3388, Relative Risk = 0.8602/0.3388 = 2.5390.

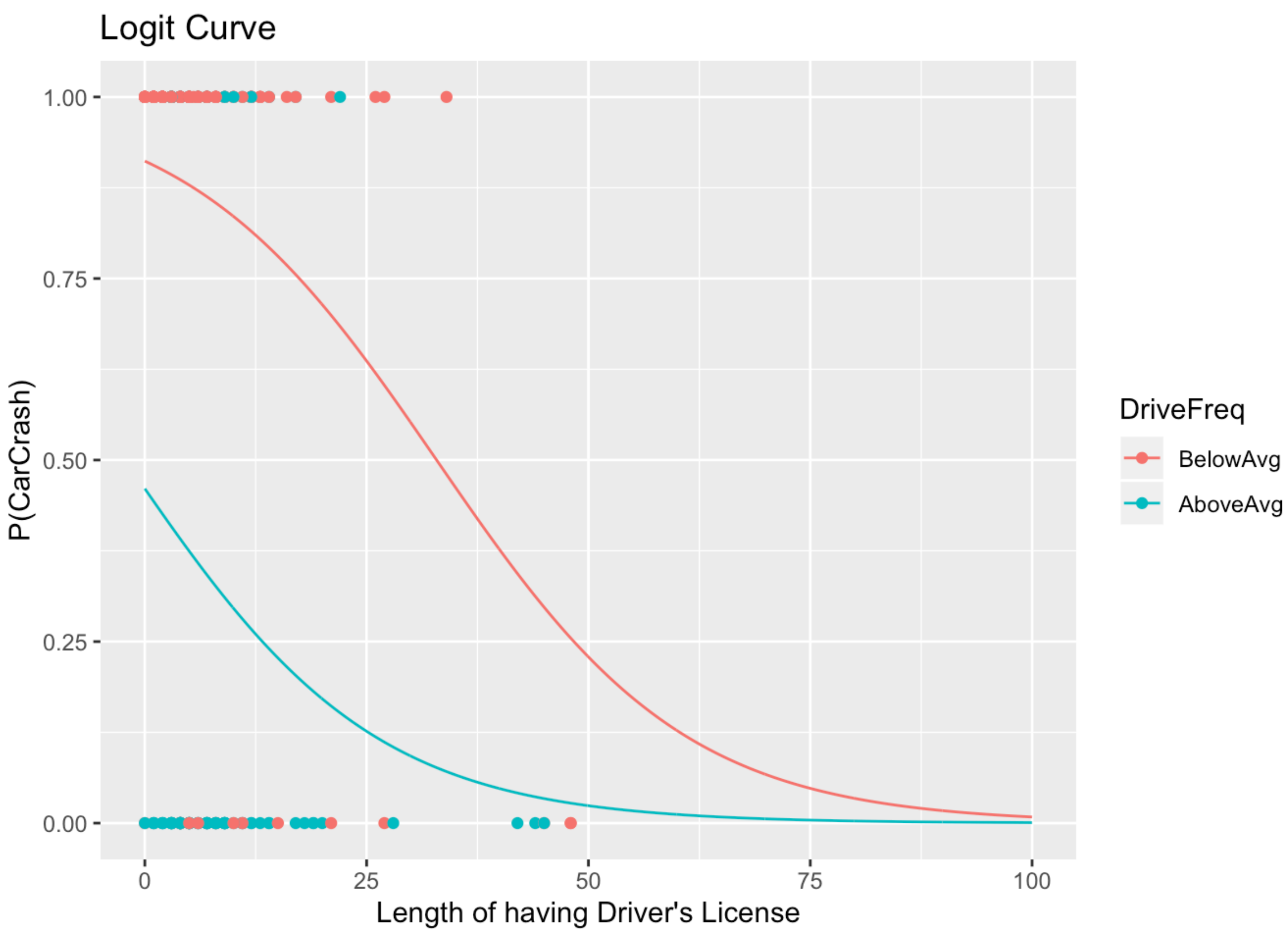
Odds(BelowAvg+Crash) = 160/26 = 6.1538, Odds(AboveAvg+Crash) = 41/80 = 0.5125, Odds Ratio = 6.1538/0.5125 = 12.0074.

Both numbers show that the below-average-frequency drivers are more crash-prone when driving.

Then I plotted the logit curve

```
generated_data <- as.data.frame(expand.grid(DL_Length = seq(min(surv
ey$DL_Length), 100, 0.1), DriveFreq = c('BelowAvg','AboveAvg')))
generated_data$probs <- plogis(predict(LB_HW5, newdata = generated_d
ata))

ggplot(generated_data, aes(x = DL_Length, y = probs, color = DriveFr
eq)) +
  geom_line() +
  geom_point(data = survey, aes(x = DL_Length, y = CarCrash)) +
  labs(x = "Length of having Driver's License", y = "P(CarCrash)", t
itle = "Logit Curve")+
  ylim(c(0, 1))
```



The logit curve shows that if you drive frequently, you are less likely to get involved in a car crash, but if you own a driver license longer, you are decreasing your odds of encountering a car crash. This result could connect the cause of car crashes to inexperience drivers who just acquire their driver's licenses more based on the result of this model.

Exercise 3

The two interaction term I added were GenderxOnCampusFreq and DL_LengthxDriveFreq.

```
LB_HW5 <- glm(CarCrash ~ Gender + DL_Length + DriveFreq + Gender*DriveFreq + OnCampusFreq*DriveFreq, data = survey, family = binomial())
summary(LB_HW5)
```

```
##
## Call:
## glm(formula = CarCrash ~ Gender + DL_Length + DriveFreq + Gender
*
##     DriveFreq + OnCampusFreq * DriveFreq, family = binomial(),
##     data = survey)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3306  -0.7875   0.4460   0.5502   1.9354
##
## Coefficients:
##                                     Estimate Std. Error z value Pr
(>|z|)
## (Intercept)                          2.71674    0.52655   5.160 2.
48e-07 ***
## GenderMale                          -0.31977    0.44401  -0.720 0.
471410
## DL_Length                           -0.06943    0.02023  -3.432 0.
000599 ***
## DriveFreqAboveAvg                   -3.11887    0.62719  -4.973 6.
60e-07 ***
## OnCampusFreqHigh                    -0.30765    0.50604  -0.608 0.
543215
## GenderMale:DriveFreqAboveAvg        -0.01212    0.60124  -0.020 0.
983919
## DriveFreqAboveAvg:OnCampusFreqHigh  1.00789    0.65583   1.537 0.
124338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.71  on 306  degrees of freedom
## Residual deviance: 287.29  on 300  degrees of freedom
## AIC: 301.29
##
## Number of Fisher Scoring iterations: 4
```

```
x <- data.frame("Variable Names" = c('Male', 'Length of having drive
rs license', 'Above average frequency of driving', 'Go on campus mor
e frequently','Male who drive frequently', 'Participants who both go
on campus and drive more frequently'), "coefficients" = LB_HW5$coeff
icients[2:7], "z-value" = c('-0.720', '-3.432', '-4.973', '-0.608',
'-0.020', '1.537'), "Significance" = c('No', 'Yes', 'Yes', 'No', 'No
', 'No'))
x
```

```
##
Variable.Names
## GenderMale
Male
## DL_Length                                                        Le
ngth of having drivers license
## DriveFreqAboveAvg                                              Abov
e average frequency of driving
## OnCampusFreqHigh
Go on campus more frequently
## GenderMale:DriveFreqAboveAvg
Male who drive frequently
## DriveFreqAboveAvg:OnCampusFreqHigh Participants who both go on ca
mpus and drive more frequently
##                                    coefficients z.value Significa
nce
## GenderMale                           -0.31976925  -0.720
No
## DL_Length                            -0.06943462  -3.432
Yes
## DriveFreqAboveAvg                    -3.11887351  -4.973
Yes
## OnCampusFreqHigh                     -0.30764853  -0.608
No
## GenderMale:DriveFreqAboveAvg         -0.01211853  -0.020
No
## DriveFreqAboveAvg:OnCampusFreqHigh    1.00788812   1.537
No
```

The AIC increases by 3.85, which means the previous model is slightly better than this one. However, if I remove some variables and leave only significant variables remain in the model, I could probably improve the model by decreasing AIC. Which is the result below, as AIC decreases by 4.69.

```
LB_HW5 <- glm(CarCrash ~ DL_Length + DriveFreq + DL_Length*DriveFreq
, data = survey, family = binomial())
summary(LB_HW5)
```

```
##
## Call:
## glm(formula = CarCrash ~ DL_Length + DriveFreq + DL_Length *
##     DriveFreq, family = binomial(), data = survey)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2826  -0.9063   0.4294   0.5379   1.6374
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   2.52845    0.31005   8.155 3.49e-16
***
## DL_Length                    -0.09546    0.02704  -3.531 0.000414
***
## DriveFreqAboveAvg            -3.00046    0.42510  -7.058 1.69e-12
***
## DL_Length:DriveFreqAboveAvg  0.06977    0.04006   1.742 0.081544
.
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.71  on 306  degrees of freedom
## Residual deviance: 288.60  on 303  degrees of freedom
## AIC: 296.6
##
## Number of Fisher Scoring iterations: 4
```