HW3 MVS

Sanjay Chhetri

11/8/2021

Part A: Logistic Regression

Instruction: "For the logistic regression analysis, you will be employing a list-wise deletion approach to handling missing data (which is also what logistic regression does by default). You will be using the following five variables in your analysis: Sex, Marital, Health, Unhappy, and Depression".

```
library(haven)
df <- read_sav("GSS2018_HW3A.sav")
head(df, 10)
```

```
## # A tibble: 10 x 5
##           Sex         Marital         Health             Unhappy Depression
##     <dbl+lbl>       <dbl+lbl>      <dbl+lbl>           <dbl+lbl>  <dbl+lbl>
##  1 1 [Male]   0 [Not married]  3 [Good]       2 [Fairly Happy]      0 [No]
##  2 0 [Female] 0 [Not married] NA             NA                    NA
##  3 1 [Male]   1 [Married]      4 [Very Good]  1 [Very Happy]        0 [No]
##  4 0 [Female] 1 [Married]      5 [Excellent] NA                     1 [Yes]
##  5 1 [Male]   0 [Not married] NA             NA                    NA
##  6 0 [Female] 0 [Not married] NA              3 [Not Very Happy]   NA
##  7 0 [Female] 0 [Not married]  3 [Good]      NA                     0 [No]
##  8 1 [Male]   0 [Not married]  4 [Very Good] NA                     0 [No]
##  9 0 [Female] 0 [Not married]  1 [Poor]       2 [Fairly Happy]      1 [Yes]
## 10 1 [Male]   1 [Married]      4 [Very Good] NA                     1 [Yes]
```

Q.N. 1

How many participants have complete data on these five variables? What percentage is this of the full sample?

```
nrow(df)
```

```
## [1] 2348
```

```
sum(complete.cases(df))
```

```
## [1] 708
```

```
df$complete_case <- complete.cases(df)
708/2348*100
```

```
## [1] 30.15332
```

Out of total 2348 observations in the datafile, 708 are complete cases. It is 30.153322 percent of the full sample.

Q.N. 2

```
test_sex <- chisq.test(table(df$Sex, df$complete_case))
test_sex
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df$Sex, df$complete_case)
## X-squared = 0.77143, df = 1, p-value = 0.3798
```

```
test_marital <- chisq.test(table(df$Marital, df$complete_case))
test_marital
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df$Marital, df$complete_case)
## X-squared = 2.4807, df = 1, p-value = 0.1153
```

```
test_Health <- chisq.test(table(df$Health, df$complete_case))
test_Health
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$Health, df$complete_case)
## X-squared = 0.22086, df = 4, p-value = 0.9943
```

```
test_unhappy <- chisq.test(table(df$Unhappy, df$Depression))
```

```
## Warning in chisq.test(table(df$Unhappy, df$Depression)): Chi-squared
## approximation may be incorrect
```

```
test_unhappy
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$Unhappy, df$Depression)
## X-squared = 26.917, df = 3, p-value = 6.128e-06
```

```
test_depression <- chisq.test(table(df$Depression, df$complete_case))
test_depression
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(df$Depression, df$complete_case)
## X-squared = 1.1308e-29, df = 1, p-value = 1
```

Sex, Marital, Health, and Depression variables each do not seem to predict the distribution of missingness in the dataset. But Unhappy variable seems to predict the distribution of missingness in the dataset.

Q.N.3

```
dfc <- df %>% filter(complete_case)
head(dfc, 10)
```

```
## # A tibble: 10 x 6
##           Sex         Marital        Health     Unhappy Depression complete_case
##     <dbl+lbl>       <dbl+lbl>     <dbl+lbl>   <dbl+lbl>  <dbl+lbl> <lgl>
##  1 1 [Male]   0 [Not married] 3 [Good]      2 [Fairly ~    0 [No]  TRUE
##  2 1 [Male]   1 [Married]     4 [Very Good] 1 [Very Ha~    0 [No]  TRUE
##  3 0 [Female] 0 [Not married] 1 [Poor]      2 [Fairly ~    1 [Yes] TRUE
##  4 1 [Male]   0 [Not married] 5 [Excellent] 1 [Very Ha~    0 [No]  TRUE
##  5 0 [Female] 0 [Not married] 4 [Very Good] 2 [Fairly ~    0 [No]  TRUE
##  6 0 [Female] 1 [Married]     5 [Excellent] 1 [Very Ha~    0 [No]  TRUE
##  7 0 [Female] 1 [Married]     3 [Good]      2 [Fairly ~    0 [No]  TRUE
##  8 1 [Male]   0 [Not married] 5 [Excellent] 1 [Very Ha~    0 [No]  TRUE
##  9 0 [Female] 1 [Married]     3 [Good]      1 [Very Ha~    0 [No]  TRUE
## 10 0 [Female] 0 [Not married] 4 [Very Good] 2 [Fairly ~    0 [No]  TRUE
```

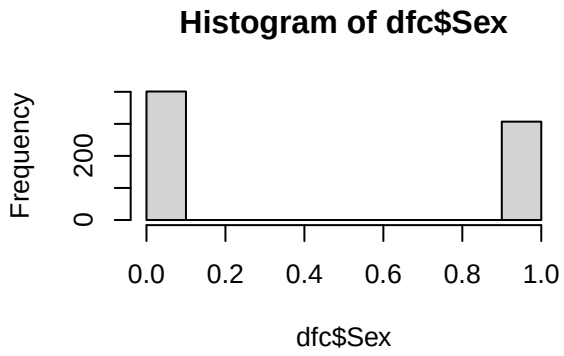
```
nrow(dfc)
```

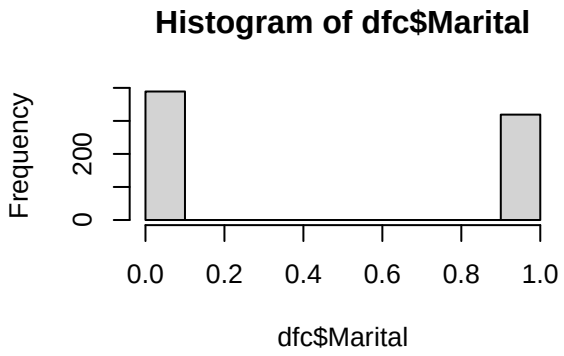
```
## [1] 708
```

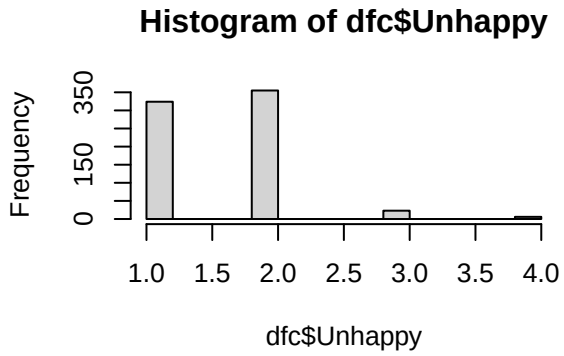
```
summary(dfc)
```

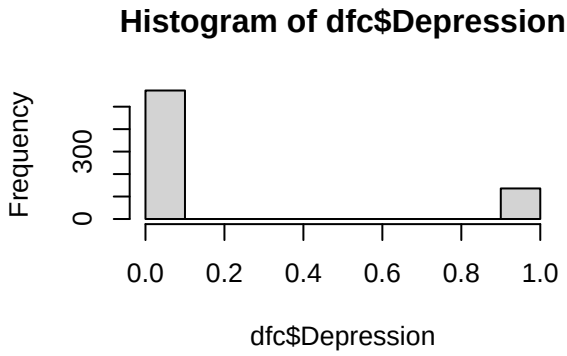
```
##       Sex            Marital           Health        Unhappy
##  Min.   :0.0000   Min.   :0.0000   Min.   :1.00   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.00   1st Qu.:1.000
##  Median :0.0000   Median :0.0000   Median :4.00   Median :2.000
##  Mean   :0.4336   Mean   :0.4506   Mean   :3.54   Mean   :1.592
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.00   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :5.00   Max.   :4.000
##    Depression     complete_case
##  Min.   :0.0000   Mode:logical
##  1st Qu.:0.0000   TRUE:708
##  Median :0.0000
##  Mean   :0.1921
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

```r
par(mfrow = c(2,2))
hist(dfc$Sex)
hist(dfc$Marital)
hist(dfc$Unhappy)
hist(dfc$Depression)
```









Q.N. 4

```r
dfc$Sex <- factor(dfc$Sex)
dfc$Marital <- factor(dfc$Marital)
logistic_model <- glm(Depression~Sex+Marital+Health+Unhappy, data = dfc, family = "binomial")
summary(logistic_model)
```

```
##
## Call:
## glm(formula = Depression ~ Sex + Marital + Health + Unhappy,
##     family = "binomial", data = dfc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5910  -0.6643  -0.5009  -0.3099   2.4739
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4229     0.5308  -0.797 0.425569
## Sex1         -1.0753     0.2235  -4.811  1.5e-06 ***
## Marital1     -0.4614     0.2157  -2.139 0.032471 *
## Health       -0.3977     0.1036  -3.839 0.000123 ***
## Unhappy       0.5381     0.1761   3.056 0.002246 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 692.76  on 707  degrees of freedom
## Residual deviance: 622.18  on 703  degrees of freedom
## AIC: 632.18
##
## Number of Fisher Scoring iterations: 5
```

```
exp(cbind(OR = coef(logistic_model), confint.default(logistic_model)))
```

```
##                    OR     2.5 %    97.5 %
## (Intercept) 0.6551270 0.2314845 1.8540823
## Sex1        0.3412075 0.2201848 0.5287494
## Marital1    0.6304248 0.4130472 0.9622033
## Health      0.6718677 0.5484153 0.8231101
## Unhappy     1.7128337 1.2128326 2.4189649
```

According to the results, sex, health, Martial and unhappy significantly predict odds of depression. Sex was significantly associated with the odds of depression. Specifically, females have 2.93 times the odds of having depression than males (95% CI: [1.89, 4.54]) Marital status was also significantly predicted the odds of depression. Specifically, unmarried people have 1.6 times the odds of being depressed than married people (95% CI: [1.03, 2.42] Health is also significantly associated with the odds of depression. For each unit worsening of health, the odds of getting depression goes up by 1.5 (95% CI: [1.21, 1.81]). State of unhappiness also predicts depression. Specifically, each unit decrease in happiness level is associated with 1.71 odds of getting depression(95% CI:[1.21, 2.41]

#Part C

```
survey <- read_sav("survey.sav")
head(survey)
```

```
## # A tibble: 6 x 104
##      id     sex   age age_group marital   child    educ  source   smoke smokenum
##   <dbl> <dbl+l> <dbl> <dbl+lbl> <dbl+l> <dbl+l> <dbl+l> <dbl+l> <dbl+l>    <dbl>
## 1   415 1 [FEM~    24 0 [young~ 4 [MAR~ 1 [YES] 5 [COM~ 7 [LIF~ 0 [NO]        NA
## 2     9 0 [MAL~    39 1 [middl~ 3 [LIV~ 1 [YES] 5 [COM~ 1 [WOR~ 1 [YES]        2
## 3   425 1 [FEM~    48 1 [middl~ 4 [MAR~ 1 [YES] 2 [SOM~ 4 [CHI~ 0 [NO]        NA
## 4   307 0 [MAL~    41 1 [middl~ 5 [REM~ 1 [YES] 2 [SOM~ 1 [WOR~ 0 [NO]         0
## 5   440 0 [MAL~    23 0 [young~ 1 [SIN~ 0 [NO]  5 [COM~ 1 [WOR~ 0 [NO]         0
## 6   484 1 [FEM~    31 1 [middl~ 4 [MAR~ 1 [YES] 5 [COM~ 7 [LIF~ 0 [NO]        NA
## # ... with 94 more variables: op1 <dbl>, Rop2 <dbl>, op3 <dbl>, Rop4 <dbl>,
## #   op5 <dbl>, Rop6 <dbl>, Rmast1 <dbl>, mast2 <dbl>, Rmast3 <dbl>,
## #   Rmast4 <dbl>, mast5 <dbl>, Rmast6 <dbl>, Rmast7 <dbl>, pn1 <dbl>,
## #   pn2 <dbl>, pn3 <dbl>, pn4 <dbl>, pn5 <dbl>, pn6 <dbl>, pn7 <dbl>,
```

```
## #   pn8 <dbl>, pn9 <dbl>, pn10 <dbl>, pn11 <dbl>, pn12 <dbl>, pn13 <dbl>,
## #   pn14 <dbl>, pn15 <dbl>, pn16 <dbl>, pn17 <dbl>, pn18 <dbl>, pn19 <dbl>,
## #   pn20 <dbl>, lifsat1 <dbl>, lifsat2 <dbl>, lifsat3 <dbl>, lifsat4 <dbl>, ...
```

```
int <- aov(Mslfest~age_group*Mlifesat, data = survey)
summary(int)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## age_group             1   2.59   2.595  11.595 0.000723 ***
## Mlifesat              1  30.71  30.708 137.228  < 2e-16 ***
## age_group:Mlifesat    1   0.88   0.876   3.914 0.048517 *
## Residuals           430  96.22   0.224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

The interaction terms is significant.

```
titanic <- read_csv("titanic.csv")
```

```
## Rows: 887 Columns: 8
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Name, Sex
## dbl (6): Survived, Pclass, Age, Siblings/Spouses Aboard, Parents/Children Ab...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

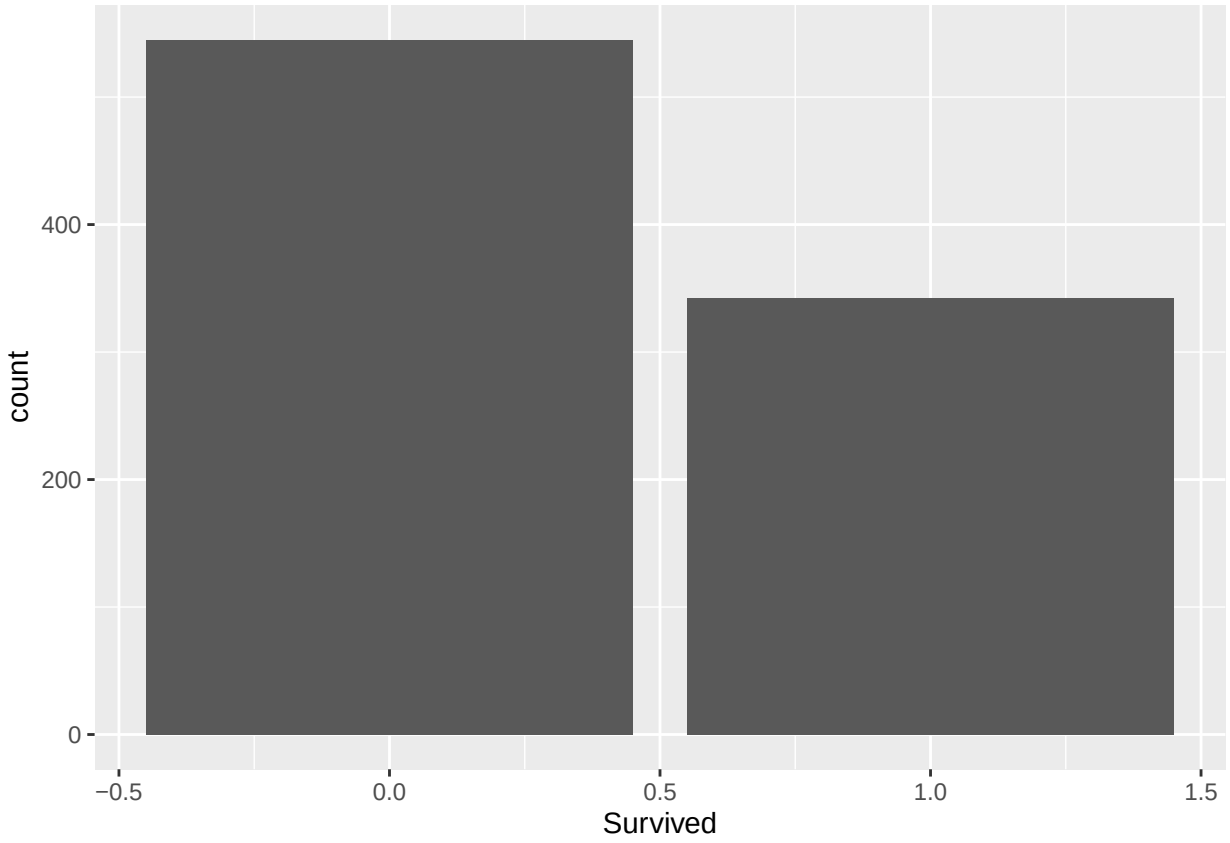
```
head(titanic)
```

```
## # A tibble: 6 x 8
##   Survived Pclass Name      Sex      Age `Siblings/Spous~ `Parents/Childr~  Fare
##      <dbl>  <dbl> <chr>     <chr>  <dbl>            <dbl>            <dbl> <dbl>
## 1        0      3 Mr. Owen~ male      22                1                0  7.25
## 2        1      1 Mrs. Joh~ female    38                1                0 71.3
## 3        1      3 Miss. La~ female    26                0                0  7.92
## 4        1      1 Mrs. Jac~ female    35                1                0 53.1
## 5        0      3 Mr. Will~ male      35                0                0  8.05
## 6        0      3 Mr. Jame~ male      27                0                0  8.46
```

```
table(titanic$Survived)
```

```
##
##   0   1
## 545 342
```

```
titanic %>% ggplot(aes(Survived))+ geom_bar()
```



Out of total 887 people, 545 people died and only 342 survived.

```
survival <- glm(Survived~factor(Pclass)+factor(Sex)+Age, data = titanic)
summary(survival)
```

```
##
## Call:
## glm(formula = Survived ~ factor(Pclass) + factor(Sex) + Age,
##     data = titanic)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.09191  -0.23675  -0.07911   0.22250   1.00040
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.101854   0.047486  23.204  < 2e-16 ***
## factor(Pclass)2 -0.190178   0.039514  -4.813 1.75e-06 ***
## factor(Pclass)3 -0.383843   0.034638 -11.082  < 2e-16 ***
## factor(Sex)male -0.494781   0.027494 -17.996  < 2e-16 ***
## Age             -0.004970   0.001005  -4.944 9.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 0.1469418)
##
##     Null deviance: 210.14  on 886  degrees of freedom
## Residual deviance: 129.60  on 882  degrees of freedom
## AIC: 823.17
##
## Number of Fisher Scoring iterations: 2
```

```
exp(cbind(OR = coef(survival), confint.default(survival)))
```

```
##                         OR     2.5 %    97.5 %
## (Intercept)     3.0097398 2.7422607 3.3033088
## factor(Pclass)2 0.8268116 0.7651956 0.8933891
## factor(Pclass)3 0.6812383 0.6365252 0.7290924
## factor(Sex)male 0.6097044 0.5777190 0.6434605
## Age             0.9950428 0.9930844 0.9970050
```

All predictors in the model had individual main effects on survival status. Passenger class (Pclass) significantly predicted survival. Specifically, compared to the first class passengers, the second class passengers had about 20% less odds of survival and the third class passengers had about 32% less odds of survival (all else being equal). Sex also predicted survival. Specifically, males had 40% less odds of survival than females. Age also predicted survival. Each year increase of age predicted 1% decrease in survival odds.