Written lesson

Author

October 17, 2020

Instructions: Using R start the data analysis using simple frequencies (number infected in your assigned county, hospitalized, died, etc.) and summary statistics (e.g., mean age and standard deviation). Then use chi-square analysis for bivariate associations (e.g., gender by death). When you use logistic regression, first use the unadjusted odds ratios then find the adjusted odds ratios.

Loading data

First load the data in R

```
Covid <- read.csv("Florida_COVID19_Case_Line_Data.csv")
head(Covid)
```

```
##    County Age   Age_group Gender Jurisdiction Travel_related Origin EDvisit
## 1 Broward  64 55-64 years Female  FL resident             No   <NA>      NO
## 2    Dade  40 35-44 years   Male  FL resident        Unknown   <NA> UNKNOWN
## 3 Broward  57 55-64 years   Male  FL resident             No   <NA> UNKNOWN
## 4 Broward  50 45-54 years Female  FL resident             No   <NA>     YES
## 5 Broward  81 75-84 years Female  FL resident             No   <NA>     YES
## 6    Dade  37 35-44 years   Male  FL resident        Unknown   <NA> UNKNOWN
##   Hospitalized Died Case_ Contact                  Case1              EventDate
## 1           NO <NA>   Yes     Yes 2020/07/19 05:00:00+00 2020/07/17 00:00:00+00
## 2      UNKNOWN <NA>   Yes    <NA> 2020/07/19 05:00:00+00 2020/07/18 23:25:02+00
## 3      UNKNOWN <NA>   Yes     Yes 2020/07/19 05:00:00+00 2020/07/18 23:24:57+00
## 4           NO <NA>   Yes     Yes 2020/07/19 05:00:00+00 2020/07/15 00:00:00+00
## 5          YES <NA>   Yes      NO 2020/07/19 05:00:00+00 2020/07/18 00:00:00+00
## 6      UNKNOWN <NA>   Yes    <NA> 2020/07/19 05:00:00+00 2020/07/18 23:25:34+00
##                ChartDate ObjectId
## 1 2020/07/19 05:00:00+00        1
## 2 2020/07/19 05:00:00+00        2
## 3 2020/07/19 05:00:00+00        3
## 4 2020/07/19 05:00:00+00        4
## 5 2020/07/19 05:00:00+00        5
## 6 2020/07/19 05:00:00+00        6
```

Filter the data, Orange COunty

Filter the data by Orange County

```r
Covid <- Covid %>%
  filter(County == "Orange") %>%
  select(County,Age,Age_group,Gender,Jurisdiction,Travel_related,Hospitalized,Died,Case_)
Covid$Age_group <- as.factor(Covid$Age_group)
Covid$Gender <- as.factor(Covid$Gender)
Covid$Jurisdiction <- as.factor(Covid$Jurisdiction)
Covid$Travel_related <- as.factor(Covid$Travel_related)
Covid$Hospitalized <- as.factor(Covid$Hospitalized)
Covid$Died <- as.factor(Covid$Died)
head(Covid)
```

```
##   County Age   Age_group Gender Jurisdiction Travel_related Hospitalized Died
## 1 Orange  54 45-54 years   Male  FL resident        Unknown      UNKNOWN <NA>
## 2 Orange  16 15-24 years Female  FL resident             No      UNKNOWN <NA>
## 3 Orange  47 45-54 years Female  FL resident             No           NO <NA>
## 4 Orange  37 35-44 years Female  FL resident        Unknown      UNKNOWN <NA>
## 5 Orange  10  5-14 years Female  FL resident        Unknown      UNKNOWN <NA>
## 6 Orange  16 15-24 years Female  FL resident        Unknown      UNKNOWN <NA>
##   Case_
## 1   Yes
## 2   Yes
## 3   Yes
## 4   Yes
## 5   Yes
## 6   Yes
```

```r
str(Covid)
```

```
## 'data.frame':    43044 obs. of  9 variables:
##  $ County        : chr  "Orange" "Orange" "Orange" "Orange" ...
##  $ Age           : int  54 16 47 37 10 16 42 67 62 20 ...
##  $ Age_group     : Factor w/ 11 levels "0-4 years","15-24 years",..: 5 2 5 4 6 2 4 8 7 2 ...
##  $ Gender        : Factor w/ 3 levels "Female","Male",..: 2 1 1 1 1 1 2 1 2 1 ...
##  $ Jurisdiction  : Factor w/ 3 levels "FL resident",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Travel_related: Factor w/ 3 levels "No","Unknown",..: 2 1 1 2 2 2 1 2 1 1 ...
##  $ Hospitalized  : Factor w/ 4 levels "","NO","UNKNOWN",..: 3 3 2 3 3 3 3 3 2 3 ...
##  $ Died          : Factor w/ 1 level "Yes": NA NA NA NA NA NA NA NA NA NA ...
##  $ Case_         : chr  "Yes" "Yes" "Yes" "Yes" ...
```

Removing NA's, UNKNOWN.

We will filter the data to remove the NA's, UNKOWN that are not useful in our data set. First we will start with Hospitalized

```r
Covid <- Covid %>% filter(Hospitalized != "" )
Covid <- Covid %>% filter(Hospitalized != "UNKNOWN")
Covid <- droplevels(Covid)
head(Covid)
```

```
##   County Age   Age_group Gender Jurisdiction Travel_related Hospitalized Died
```

```
## 1 Orange  47 45-54 years Female  FL resident             No            NO <NA>
## 2 Orange  62 55-64 years   Male  FL resident             No            NO <NA>
## 3 Orange  16 15-24 years   Male  FL resident             No            NO <NA>
## 4 Orange  16 15-24 years   Male  FL resident             No            NO <NA>
## 5 Orange  19 15-24 years   Male  FL resident             No            NO <NA>
## 6 Orange  51 45-54 years   Male  FL resident             No            NO <NA>
##   Case_
## 1   Yes
## 2   Yes
## 3   Yes
## 4   Yes
## 5   Yes
## 6   Yes
```

```r
str(Covid)
```

```
## 'data.frame':    14732 obs. of  9 variables:
##  $ County         : chr  "Orange" "Orange" "Orange" "Orange" ...
##  $ Age            : int  47 62 16 16 19 51 41 1 28 62 ...
##  $ Age_group      : Factor w/ 11 levels "0-4 years","15-24 years",..: 5 7 2 2 2 5 4 1 3 7 ...
##  $ Gender         : Factor w/ 3 levels "Female","Male",..: 1 2 2 2 2 2 1 2 1 1 ...
##  $ Jurisdiction   : Factor w/ 3 levels "FL resident",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Travel_related: Factor w/ 3 levels "No","Unknown",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ Hospitalized   : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Died           : Factor w/ 1 level "Yes": NA NA NA NA NA NA NA NA NA 1 ...
##  $ Case_          : chr  "Yes" "Yes" "Yes" "Yes" ...
```

Now we will continue removing the UNKOWN from Travel_related, Gender, and Age_group

```r
Covid <- Covid %>% filter(Age_group != "Unknown")
Covid <- Covid %>% filter(Travel_related != "Unknown")
Covid <- Covid %>% filter(Gender != "Unknown")
Covid <- droplevels(Covid)
str(Covid)
```

```
## 'data.frame':    12964 obs. of  9 variables:
##  $ County         : chr  "Orange" "Orange" "Orange" "Orange" ...
##  $ Age            : int  47 62 16 16 19 51 41 1 28 73 ...
##  $ Age_group      : Factor w/ 10 levels "0-4 years","15-24 years",..: 5 7 2 2 2 5 4 1 3 8 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 1 2 1 2 ...
##  $ Jurisdiction   : Factor w/ 3 levels "FL resident",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Travel_related: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Hospitalized   : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Died           : Factor w/ 1 level "Yes": NA NA NA NA NA NA NA NA NA NA ...
##  $ Case_          : chr  "Yes" "Yes" "Yes" "Yes" ...
```

Frequency tables and summary statistics

Now we can start with the frequecy tables.

Some tables between categorical variables The table below show that we have some missing values for Hospitalized, and we can remove them but its up to you to remove those NA's in the variables, I do recommend because we are just interested in those who answered something meaninful, we might loose a lot of information about variables and observations but we will have more accurate results. We can see that the majority of observations didnt go to a hospital, and the most observations fall between 15-24 years and 25-34 years.

```
table(Covid$Age_group, Covid$Hospitalized)
```

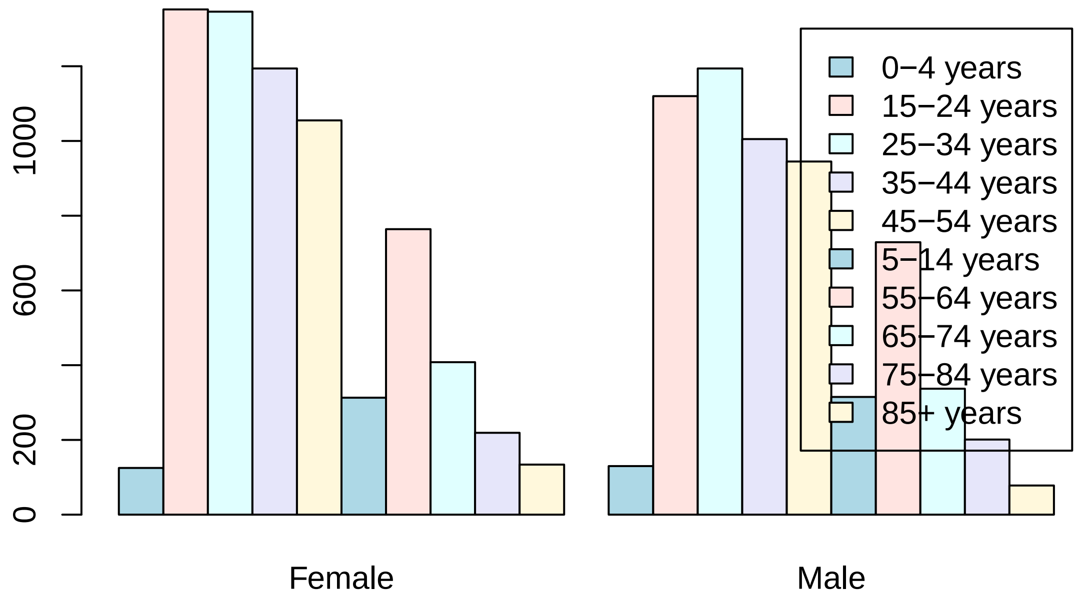
```
##
##                NO  YES
##   0-4 years   248    7
##   15-24 years 2436   36
##   25-34 years 2448   92
##   35-44 years 2047  152
##   45-54 years 1813  187
##   5-14 years   621    7
##   55-64 years 1214  279
##   65-74 years  514  231
##   75-84 years  213  207
##   85+ years     78  134
```

In this table we can see that the observations for female and male are similar for each age group category. We can also make a bar plot so you can see how it looks the table.

```
table1 <- table(Covid$Age_group, Covid$Gender)
table1
```

```
##
##              Female Male
##   0-4 years      125  130
##   15-24 years   1352 1120
##   25-34 years   1346 1194
##   35-44 years   1194 1005
##   45-54 years   1055  945
##   5-14 years     313  315
##   55-64 years    764  729
##   65-74 years    408  337
##   75-84 years    219  201
##   85+ years      134   78
```

```
barplot(table1,
         col = c("lightblue", "mistyrose", "lightcyan",
                 "lavender", "cornsilk"),
        legend = rownames(table1), beside=TRUE)
```



For this part the majority of observations are in the FL resident group, no matter the age group they are.

```
table(Covid$Age_group, Covid$Jurisdiction)
```

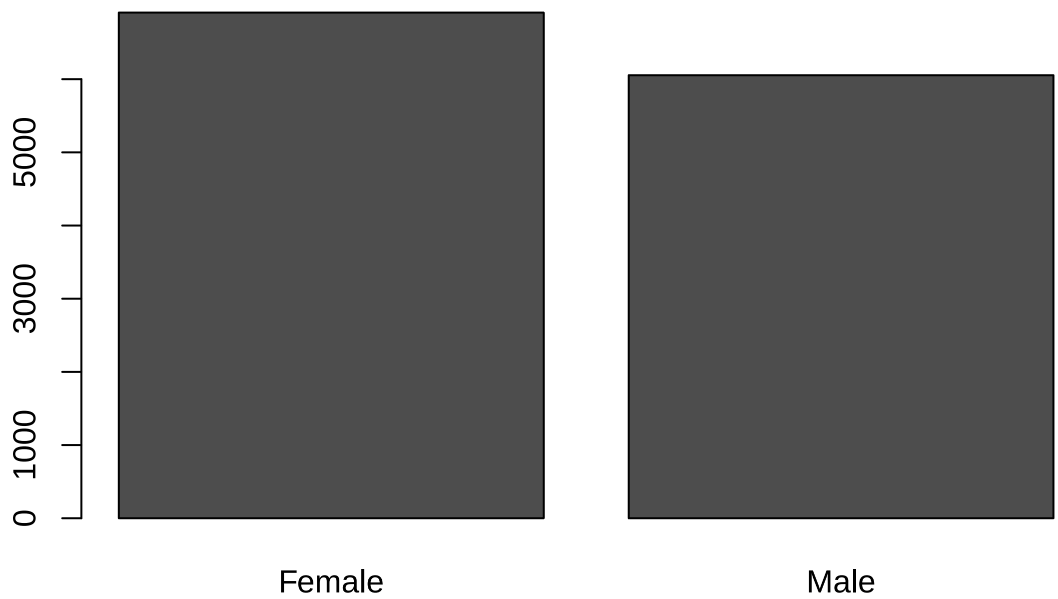
```
##
##                FL resident Non-FL resident Not diagnosed/isolated in FL
##   0-4 years            254               1                            0
##   15-24 years         2455              17                            0
##   25-34 years         2529              11                            0
##   35-44 years         2189              10                            0
##   45-54 years         1988              12                            0
##   5-14 years           628               0                            0
##   55-64 years         1484               9                            0
##   65-74 years          730              14                            1
##   75-84 years          409              11                            0
##   85+ years            211               1                            0
```

All the cases in the data are positive and the majority of positive covid-19 cases are females but the difference is not huge, its a small difference.

```
table(Covid$Case_, Covid$Gender)
```

```
##
##       Female Male
##   Yes   6910 6054
```

```r
barplot(table(Covid$Case_, Covid$Gender))
```



The majority of positive cases of covid-19 are between 15-24 years, 25-34 years and 35-44 years.

```r
table(Covid$Age_group, Covid$Case_)
```

```
##
##                 Yes
##   0-4 years     255
##   15-24 years  2472
##   25-34 years  2540
##   35-44 years  2199
##   45-54 years  2000
##   5-14 years    628
##   55-64 years  1493
##   65-74 years   745
##   75-84 years   420
##   85+ years     212
```

Summary statistics for Age The median age is 37 which means that 50% of the observations have an age below 36 and 50% of the observations have an age above 36.

```r
summary(Covid$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   24.00   37.00   39.26   52.00  106.00
```

Chi square test

Chi square for bivariate analysis Now by doing the chi square test we can see if 2 categorical variables are related or not, by using chi square for gender and died, we can see that p value is smaller than alpha 0.05 so we can conclude that the gender and died variables are dependent, that means one of the 2 gender is more likely to die.

```r
table(Covid$Gender, Covid$Died)
```

```
##
##          Yes
##   Female 136
##   Male   188
```

```r
chisq.test(table(Covid$Gender, Covid$Died))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(Covid$Gender, Covid$Died)
## X-squared = 8.3457, df = 1, p-value = 0.003866
```

By doing the chi square test between age group and hospitalization, the p value is smaller than alpha 0.05 so we can say that age and hospitalization are related, in other words depending in your age it is more likely to be hospitalized.

```r
table(Covid$Age_group, Covid$Hospitalized)
```

```
##
##                 NO  YES
##   0-4 years    248    7
##   15-24 years 2436   36
##   25-34 years 2448   92
##   35-44 years 2047  152
##   45-54 years 1813  187
##   5-14 years   621    7
##   55-64 years 1214  279
##   65-74 years  514  231
##   75-84 years  213  207
##   85+ years     78  134
```

```r
chisq.test(table(Covid$Age_group, Covid$Hospitalized))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Covid$Age_group, Covid$Hospitalized)
## X-squared = 2231.7, df = 9, p-value < 2.2e-16
```

Here we did a chi square test between age and gender, where we found a p value lower than alpha 0.05, so age and gender are related.

```r
table(Covid$Age_group, Covid$Gender)
```

```
##
##               Female Male
##   0-4 years      125  130
##   15-24 years   1352 1120
##   25-34 years   1346 1194
##   35-44 years   1194 1005
##   45-54 years   1055  945
##   5-14 years     313  315
##   55-64 years    764  729
##   65-74 years    408  337
##   75-84 years    219  201
##   85+ years      134   78
```

```r
chisq.test(table(Covid$Age_group, Covid$Gender))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Covid$Age_group, Covid$Gender)
## X-squared = 19.985, df = 9, p-value = 0.018
```

Here the p value is lower than 0.05, so we can say that Travel related depends on Hospitalized.

```r
table(Covid$Travel_related, Covid$Hospitalized)
```

```
##
##          NO   YES
##   No  11033  1230
##   Yes   599   102
```

```r
chisq.test(table(Covid$Travel_related, Covid$Hospitalized))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(Covid$Travel_related, Covid$Hospitalized)
## X-squared = 14.212, df = 1, p-value = 0.0001633
```

In this chi square the p value is smaller than alpha 0.05 so we can conclude that Jurisdiction and age are related.

```r
table(Covid$Age_group, Covid$Jurisdiction)
```

```
##
##               FL resident Non-FL resident Not diagnosed/isolated in FL
##   0-4 years           254               1                            0
##   15-24 years        2455              17                            0
##   25-34 years        2529              11                            0
```

```
##   35-44 years       2189            10                           0
##   45-54 years       1988            12                           0
##   5-14 years         628             0                           0
##   55-64 years       1484             9                           0
##   65-74 years        730            14                           1
##   75-84 years        409            11                           0
##   85+ years          211             1                           0
```

```
chisq.test(table(Covid$Age_group, Covid$Jurisdiction))
```

```
## Warning in chisq.test(table(Covid$Age_group, Covid$Jurisdiction)): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Covid$Age_group, Covid$Jurisdiction)
## X-squared = 65.836, df = 18, p-value = 2.258e-07
```

Logistic Regression

Making the logistic regression… Lets say you want to predict the probability that you will be hospitalized based on gender, age , Jurisdiction and Travel related

```
logic <- glm(formula=Hospitalized ~ Gender+Age+Jurisdiction+Travel_related, data=Covid, family = binomi

summary(logic)
```

```
##
## Call:
## glm(formula = Hospitalized ~ Gender + Age + Jurisdiction + Travel_related,
##     family = binomial, data = Covid)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8802  -0.4488  -0.2747  -0.1844   3.3170
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -5.497316   0.111697 -49.216  < 2e-16
## GenderMale                               0.319895   0.064058   4.994 5.92e-07
## Age                                      0.065048   0.001784  36.460  < 2e-16
## JurisdictionNon-FL resident              1.656781   0.280018   5.917 3.28e-09
## JurisdictionNot diagnosed/isolated in FL 12.482674 196.967726   0.063    0.949
## Travel_relatedYes                        0.157407   0.133104   1.183    0.237
##
## (Intercept)                             ***
## GenderMale                              ***
## Age                                     ***
## JurisdictionNon-FL resident             ***
## JurisdictionNot diagnosed/isolated in FL
```

```
## Travel_relatedYes
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8584.1  on 12963  degrees of freedom
## Residual deviance: 6805.6  on 12958  degrees of freedom
## AIC: 6817.6
##
## Number of Fisher Scoring iterations: 10
```

By looking the p values (Pr > z) you can see that all the variables are significant, but JurisdictionNont diagnosed/isolated in FL, and Travel_relatedYes have p values greater than 0.05, so we will say that these 2 variables are not significant in the model.

The model to predict the probability of being hospitalized is:

$$P(Hospitalized) = \frac{e^{-5.497+0.319(Male)+0.065(Age)+1.657(JurisdictionNon-Flresident)+12.483(JurisdictionNotdiagnosed/isolatedinFL)}}{1+e^{-5.497+0.319(Male)+0.065(Age)+1.657(JurisdictionNon-Flresident)+12.483(JurisdictionNotdiagnosed/isolatedinFL)}}$$

We can check how our model can predict accuracy the probability of being hospitalized

```r
hosp_predict <- predict(logic, newdata=Covid, type="response")
length(hosp_predict)
```

```
## [1] 12964
```

```r
head(hosp_predict)
```

```
##          1          2          3          4          5          6
## 0.08016974 0.24150939 0.01572511 0.01572511 0.01904912 0.13470841
```

We will recode the prediction variable, if the probability is greater than 0.5 then it will be hospitalized, if not then it will not be hospitalized

```r
hosp_predict <- ifelse(hosp_predict < 0.5, "NO","YES")
head(hosp_predict)
```

```
##    1    2    3    4    5    6
## "NO" "NO" "NO" "NO" "NO" "NO"
```

Now we will make the confusion matrix that is the predicted hospitalized and the observed hospitalized

```r
confusionmatrix <- table(Covid$Hospitalized, hosp_predict)
confusionmatrix
```

```
##      hosp_predict
##          NO   YES
##   NO  11504   128
##   YES  1120   212
```

The misclassification are those where you predict No but the Observed is Yes, those are 1120, and those where you predict Yes but the Observed is No, 128 divided by the total.

```r
misclass <- (1120+128)/length(hosp_predict)
misclass
```

```
## [1] 0.09626658
```

Adjusted Odds Ratio Before we calculated the model with unadjusted odds ratio, now we can make the logistic regression model using adjusted Odds Ratio Unadjusted Odds Ratio

```r
logic$coefficients
```

```
##                               (Intercept)
##                               -5.49731596
##                                GenderMale
##                                0.31989547
##                                       Age
##                                0.06504836
##                  JurisdictionNon-FL resident
##                                1.65678136
## JurisdictionNot diagnosed/isolated in FL
##                               12.48267371
##                          Travel_relatedYes
##                                0.15740718
```

Adjusted Odds Ratio

```r
or <- exp(logic$coefficients)
round(or,2)
```

```
##                               (Intercept)
##                                      0.00
##                                GenderMale
##                                      1.38
##                                       Age
##                                      1.07
##                  JurisdictionNon-FL resident
##                                      5.24
## JurisdictionNot diagnosed/isolated in FL
##                                 263728.04
##                          Travel_relatedYes
##                                      1.17
```