Healthcare and Medical Analytics

Individual Assignment

Konstantinos Paganopoulos

First, we load all the libraries that we need.

After that, we load the data that we need for our model:

```r
data_model = read.csv("data_model.csv")
head(data_model)
```

```
##   X allergies gender_Male stress_Yes arrested_Yes rape_Yes nutrition_normal
## 1 0        No           0          0            0        0                0
## 2 1        No           1          0            1        1                0
## 3 2        No           1          0            1        0                1
## 4 3        No           1          0            0        0                1
## 5 4        No           1          0            1        0                1
## 6 5        No           1          0            1        0                0
##   nutrition_unhealthy assets_poor assets_wealthy ethnicity_black
## 1                   1           0              0               1
## 2                   1           1              0               1
## 3                   0           0              0               0
## 4                   0           0              0               0
## 5                   0           1              0               1
## 6                   1           0              0               0
##   ethnicity_indian ethnicity_white
## 1                0               0
## 2                0               0
## 3                0               1
## 4                0               1
## 5                0               0
## 6                0               1
```

Let's now build several logistic regression models to examine how each factor affects the relationship between gender and seasonal allergies.

First, we start by examining the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male, family = binomial(link = 'logit'), data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male, family = binomial(link = "logit"),
##     data = data_model)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6583  -0.6583  -0.5593  -0.5593   1.9660
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.41902    0.05081 -27.926  < 2e-16 ***
## gender_Male -0.35722    0.08026  -4.451 8.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4166.1  on 4560  degrees of freedom
## AIC: 4170.1
##
## Number of Fisher Scoring iterations: 4
```

We continue by examining how stress affects the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male + stress_Yes, family = binomial(link = 'logit'),
             data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + stress_Yes, family = binomial(link = "logit"),
##     data = data_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8628  -0.6493  -0.5544  -0.5544   1.9743
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.44953    0.05190 -27.930  < 2e-16 ***
## gender_Male -0.34565    0.08044  -4.297 1.73e-05 ***
## stress_Yes   0.65305    0.18787   3.476 0.000509 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4155.1  on 4559  degrees of freedom
## AIC: 4161.1
##
## Number of Fisher Scoring iterations: 4
```

We continue by examining how being arrested affects the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male + arrested_Yes, family = binomial(link = 'logit'),
             data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + arrested_Yes, family = binomial(link = "logit"),
##     data = data_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6657  -0.6657  -0.5754  -0.5363   2.0052
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.39410    0.05294 -26.334  < 2e-16 ***
## gender_Male  -0.32068    0.08324  -3.852 0.000117 ***
## arrested_Yes -0.15194    0.09405  -1.616 0.106200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4163.5  on 4559  degrees of freedom
## AIC: 4169.5
##
## Number of Fisher Scoring iterations: 4
```

We continue by examining how being raped affects the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male + rape_Yes, family = binomial(link = 'logit'),
             data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + rape_Yes, family = binomial(link = "logit"),
##     data = data_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7417  -0.6354  -0.5559  -0.5559   1.9717
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.49744    0.05760 -25.997  < 2e-16 ***
## gender_Male -0.29185    0.08337  -3.501 0.000464 ***
## rape_Yes     0.34745    0.11082   3.135 0.001718 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4156.6  on 4559  degrees of freedom
## AIC: 4162.6
##
## Number of Fisher Scoring iterations: 4
```

We continue by examining how normal nutrition affects the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male + nutrition_normal,
             family = binomial(link = 'logit'),data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + nutrition_normal, family = binomial(link = "logit"),
##     data = data_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6735  -0.6455  -0.5738  -0.5493   1.9829
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.46246    0.06286 -23.266  < 2e-16 ***
## gender_Male      -0.35263    0.08036  -4.388 1.14e-05 ***
## nutrition_normal  0.09446    0.07919   1.193    0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4164.7  on 4559  degrees of freedom
## AIC: 4170.7
##
## Number of Fisher Scoring iterations: 4
```

We continue by examining how unhealthy nutrition affects the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male + nutrition_unhealthy,
             family = binomial(link = 'logit'), data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + nutrition_unhealthy,
##     family = binomial(link = "logit"), data = data_model)
##
```

```
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6632  -0.6632  -0.5655  -0.5497   1.9822  
## 
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         -1.40242    0.05568 -25.185  < 2e-16 ***
## gender_Male         -0.34972    0.08092  -4.322 1.55e-05 ***
## nutrition_unhealthy -0.06137    0.08536  -0.719    0.472    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4165.6  on 4559  degrees of freedom
## AIC: 4171.6
## 
## Number of Fisher Scoring iterations: 4
```

We continue by examining how being poor affects the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male + assets_poor, family = binomial(link = 'logit'),
             data_model)
summary(model)
```

```
## 
## Call:
## glm(formula = allergies ~ gender_Male + assets_poor, family = binomial(link = "logit"), 
##     data = data_model)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.6732  -0.6732  -0.5683  -0.5311   2.0141  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -1.36916    0.05848 -23.411  < 2e-16 ***
## gender_Male -0.37224    0.08077  -4.608 4.06e-06 ***
## assets_poor -0.14586    0.08718  -1.673   0.0943 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4163.3  on 4559  degrees of freedom
## AIC: 4169.3
## 
## Number of Fisher Scoring iterations: 4
```

We continue by examining how being wealthy affects the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male + assets_wealthy,
             family = binomial(link = 'logit'), data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + assets_wealthy, family = binomial(link = "logit"),
##     data = data_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6745  -0.6514  -0.5715  -0.5513   1.9794
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.44236    0.05685 -25.373  < 2e-16 ***
## gender_Male    -0.36462    0.08067  -4.520 6.19e-06 ***
## assets_wealthy  0.07745    0.08315   0.931    0.352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4165.2  on 4559  degrees of freedom
## AIC: 4171.2
##
## Number of Fisher Scoring iterations: 4
```

We continue by examining how black ethnicity affects the relationship between gender and seasonal allergies:

```r
model <- glm(allergies ~ gender_Male + ethnicity_black,
             family = binomial(link = 'logit'), data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + ethnicity_black, family = binomial(link = "logit"),
##     data = data_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6784  -0.6784  -0.5737  -0.5038   2.0626
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.35195    0.05535 -24.428  < 2e-16 ***
## gender_Male     -0.36909    0.08043  -4.589 4.46e-06 ***
## ethnicity_black -0.27917    0.09726  -2.870   0.0041 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4157.5  on 4559  degrees of freedom
## AIC: 4163.5
##
## Number of Fisher Scoring iterations: 4
```

We continue by examining how indian ethnicity affects the relationship between gender and seasonal allergies:

```
model <- glm(allergies ~ gender_Male + ethnicity_indian,
             family = binomial(link = 'logit'), data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + ethnicity_indian, family = binomial(link = "logit"),
##     data = data_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8398  -0.6569  -0.5579  -0.5579   1.9683
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.42382    0.05097 -27.935  < 2e-16 ***
## gender_Male      -0.35762    0.08028  -4.455 8.39e-06 ***
## ethnicity_indian  0.56294    0.39192   1.436    0.151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4164.2  on 4559  degrees of freedom
## AIC: 4170.2
##
## Number of Fisher Scoring iterations: 4
```

Now we combine all of the above factors and check how all of them affect the relationship between gender and seasonal allergies:

```
model <- glm(allergies ~ gender_Male + stress_Yes + arrested_Yes + rape_Yes
             + nutrition_normal+ nutrition_unhealthy + assets_poor
             + assets_wealthy + ethnicity_black + ethnicity_indian
             + ethnicity_white, family = binomial(link = 'logit'), data_model)
summary(model)
```

```
##
## Call:
## glm(formula = allergies ~ gender_Male + stress_Yes + arrested_Yes +
##     rape_Yes + nutrition_normal + nutrition_unhealthy + assets_poor +
```

```
##     assets_wealthy + ethnicity_black + ethnicity_indian + ethnicity_white,
##     family = binomial(link = "logit"), data = data_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1788  -0.6468  -0.5796  -0.5125   2.1663
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.86671    0.27280  -6.843 7.76e-12 ***
## gender_Male          -0.27125    0.08865  -3.060  0.00221 **
## stress_Yes            0.56687    0.19275   2.941  0.00327 **
## arrested_Yes         -0.17747    0.09602  -1.848  0.06458 .
## rape_Yes              0.30564    0.11381   2.686  0.00724 **
## nutrition_normal      0.12627    0.10093   1.251  0.21091
## nutrition_unhealthy  0.06957    0.11026   0.631  0.52806
## assets_poor          -0.11955    0.09928  -1.204  0.22852
## assets_wealthy        0.01774    0.09459   0.188  0.85127
## ethnicity_black       0.16773    0.26980   0.622  0.53415
## ethnicity_indian      0.92787    0.46982   1.975  0.04827 *
## ethnicity_white       0.40188    0.25781   1.559  0.11904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4186.2  on 4561  degrees of freedom
## Residual deviance: 4130.1  on 4550  degrees of freedom
## AIC: 4154.1
##
## Number of Fisher Scoring iterations: 4
```

From the above results, we can clearly see that the gender of someone, the fact that she/he has been raped or not, as well as if the respondent has stress, are the three most significant factors that determine the frequency of his/her seasonal allergies occurence. The later was expected, since women are more likely to be abused and raped, and because according to our literature review the proportion of women feeling stressed is higher than that of men.

Moreover, indian ethnicity seems to play a role in determining the frequency of seasonal allergies. As for the variable arrested or not, we could say that is quite significant, probably because it is associated with gender as men are more likely to be arrested due to being more violent.

Other socioeconomic factors, such as total assets that somebody owns, black or white ethnicity, or nutrition, even if that is unhealthy, don't seem to affect at all the frequency of seasonal allergies between the two genders.