hw6

Jared Yu

November 26, 2018

```
library(readxl)
flu <- read_excel('flu.xlsx')
flu
```

```
## # A tibble: 159 x 6
##     Shot   Age `Health Awareness` Gender X__1  X__2
##    <dbl> <dbl>              <dbl>  <dbl> <lgl> <chr>
##  1    0.   59.                52.     0. NA    <NA>
##  2    0.   61.                55.     1. NA    Gender: 0=female, 1=male
##  3    1.   82.                51.     0. NA    Shot: 0=no flu shot, 1=rec~
##  4    0.   51.                70.     0. NA    <NA>
##  5    0.   53.                70.     0. NA    <NA>
##  6    0.   62.                49.     1. NA    <NA>
##  7    0.   51.                69.     1. NA    <NA>
##  8    0.   70.                54.     1. NA    <NA>
##  9    0.   71.                65.     1. NA    <NA>
## 10    0.   55.                58.     1. NA    <NA>
## # ... with 149 more rows
```

```
flu$X__1 = NULL
flu$X__2 = NULL
flu.fit = glm(Shot~., family=binomial(), data=flu)
flu.fit
```

```
##
## Call:  glm(formula = Shot ~ ., family = binomial(), data = flu)
##
## Coefficients:
##        (Intercept)                 Age  `Health Awareness`
##           -1.17716             0.07279            -0.09899
##             Gender
##            0.43397
##
## Degrees of Freedom: 158 Total (i.e. Null);  155 Residual
## Null Deviance:       134.9
## Residual Deviance: 105.1     AIC: 113.1
```

1 a) The MLE for β_0 is:

```
flu.fit$coefficients[1]
```

```
## (Intercept)
##   -1.177159
```

The MLE for β_1 is:

```
flu.fit$coefficients[2]
```

```
##        Age
## 0.07278802
```

The MLE for β_2 is:

```
flu.fit$coefficients[3]
```

```
## `Health Awareness`
##        -0.09898649
```

The MLE for β_3 is:

```
flu.fit$coefficients[4]
```

```
##    Gender
## 0.4339749
```

Below are the standard errors for the different coeffcieints: $\beta_0, \beta_1, \beta_2,$ and β_3 :

```
summary(flu.fit)$coefficients[,2]
```

```
##        (Intercept)                Age `Health Awareness`
##         2.98242265         0.03038087         0.03347856
##             Gender
##         0.52179407
```

$s(\beta_0) = 2.98242265$, $s(\beta_1) = 0.03038087$, $s(\beta_2) = 0.03347856$, $s(\beta_3) = 0.52179407$

The estimated logistic regression function is therefore: $log(\frac{\pi}{1-\pi}) = -1.177159 + 0.07278802X_1 - 0.09898649X_2 + 0.43397485X_3$ Where π is probability of getting a flu shot, X_1 is the age, X_2 is the Health Awareness, and X_3 is the Gender.

1 b) $exp(\hat{\beta_1})$

```
exp(flu.fit$coefficients[2])
```

```
##      Age
## 1.075503
```

The odds of getting a flu shot at Age $X + 1$ is about 1.075503 times the odds of getting the flu at Age X .

$exp(\hat{\beta}_2)$

```
exp(flu.fit$coefficients[3])
```

```
## `Health Awareness`
##          0.9057549
```

The odds of getting a flu shot at Health Awareness $X + 1$ is about 0.9057549 times the odds of getting the flu at Health Awareness X .

$exp(\hat{\beta}_3)$

```
exp(flu.fit$coefficients[4])
```

```
##  Gender
## 1.54338
```

The odds of getting a flu shot at Gender $X + 1$ is about 1.54338 times the odds of getting the flu at Gender X .

1 c)

```
test.55 = data.frame(t(c(55, 60, 1)))
colnames(test.55) = c('Age', 'Health Awareness', 'Gender')
probit=function(x){
  return(exp(x)/(1+exp(x)))
}
predict(flu.fit, test.55, type='response')
```

```
##          1
## 0.06422197
```

The probability that a randomly chosen 55 year old male with health awareness index 60 will get a flu shot is 0.06422197.

```
result=predict(flu.fit, test.55,se.fit = TRUE)
result$fit
```

```
##         1
## -2.679033
```

```
c(probit(result$fit-1.96*result$se.fit), probit(result$fit+1.96*result$se.fit))
```

```
##          1          1
## 0.02470225 0.15680208
```

The confidence interval for a randomly chosen 55 year old male with health awareness index 60 getting a flu shot is $(0.02470225, 0.15680208)$ at confidence level $\alpha = 0.05$.

1 d) $H_0 : \beta_3 = 0$, $H_1 : \beta_3 \neq 0$

```
flu.fit2 = glm(Shot~Age + get('Health Awareness'), family=binomial(), data=flu)
flu.fit
```

```
## 
## Call:  glm(formula = Shot ~ ., family = binomial(), data = flu)
## 
## Coefficients:
##        (Intercept)                 Age  `Health Awareness`  
##           -1.17716             0.07279            -0.09899  
##             Gender  
##            0.43397  
## 
## Degrees of Freedom: 158 Total (i.e. Null);  155 Residual
## Null Deviance:       134.9 
## Residual Deviance: 105.1     AIC: 113.1
```

```
flu.fit2
```

```
##
## Call:  glm(formula = Shot ~ Age + get("Health Awareness"), family = binomial(),
##     data = flu)
##
## Coefficients:
##             (Intercept)                      Age  get("Health Awareness")
##                -1.45778                  0.07787                 -0.09547
##
## Degrees of Freedom: 158 Total (i.e. Null);  156 Residual
## Null Deviance:       134.9
## Residual Deviance: 105.8     AIC: 111.8
```

```
test_stat=105.8-105.1
test_stat>qchisq(0.95,1)
```

```
## [1] FALSE
```

```
1-pchisq(test_stat, 1)
```

```
## [1] 0.4027837
```

The p-value is 0.4027837, and so we fail to reject the null hypothesis that $\beta_3 = 0$ at $\alpha = 0.05$.

2 a) Let the model be $\pi' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$

β_1 represents Age, β_2 represents Health Awareness, β_3 represents Age^2, β_4 represents Health Awareness^2, and β_5 represents Age * Health Awareness.

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ against $H_1 : H_0 \text{ is false}$

```
fit3 = glm(Shot~Age + get('Health Awareness') + I(Age^2) + I(get('Health Awareness')^
2) + Age*get('Health Awareness'), family=binomial(), data=flu)
fit3
```

```
##
## Call:  glm(formula = Shot ~ Age + get("Health Awareness") + I(Age^2) +
##     I(get("Health Awareness")^2) + Age * get("Health Awareness"),
##     family = binomial(), data = flu)
##
## Coefficients:
##                  (Intercept)                           Age
##                   13.3727995                     0.0348349
##      get("Health Awareness")                      I(Age^2)
##                   -0.6026948                    -0.0006755
## I(get("Health Awareness")^2)   Age:get("Health Awareness")
##                    0.0031696                     0.0025201
##
## Degrees of Freedom: 158 Total (i.e. Null);  153 Residual
## Null Deviance:       134.9
## Residual Deviance: 104.3     AIC: 116.3
```

```
test_stat=105.8-104.3
test_stat>qchisq(0.95,3)
```

```
## [1] FALSE
```

The p-value for the test is:

```
1-pchisq(test_stat, 3)
```

```
## [1] 0.6822703
```

Therefore we fail to reject the null hypothesis that the second order terms previously mentioned are equal to 0 at $\alpha = 0.05$.

2 b)

```
library(MASS)
step <- stepAIC(fit3, direction='backward', trace=FALSE)
step$anova
```

```
## Stepwise Model Path 
## Analysis of Deviance Table
## 
## Initial Model:
## Shot ~ Age + get("Health Awareness") + I(Age^2) + I(get("Health Awareness")^2) + 
##     Age * get("Health Awareness")
## 
## Final Model:
## Shot ~ Age + get("Health Awareness")
## 
## 
##                              Step Df   Deviance Resid. Df Resid. Dev
## 1                                                     153   104.2614
## 2                     - I(Age^2)  1 0.04897342       154   104.3104
## 3  - Age:get("Health Awareness")  1 0.53437486       155   104.8448
## 4 - I(get("Health Awareness")^2)  1 0.95060581       156   105.7954
##        AIC
## 1 116.2614
## 2 114.3104
## 3 112.8448
## 4 111.7954
```

```
summary(step)
```

```
##
## Call:
## glm(formula = Shot ~ Age + get("Health Awareness"), family = binomial(),
##     data = flu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4479  -0.5708  -0.3390  -0.1629   2.8430
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.45778    2.91534  -0.500  0.61705
## Age                      0.07787    0.02970   2.622  0.00873 **
## get("Health Awareness") -0.09547    0.03241  -2.946  0.00322 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.80  on 156  degrees of freedom
## AIC: 111.8
##
## Number of Fisher Scoring iterations: 6
```

So the estimated logistic regression model is written as:

$\hat{\pi}' = -1.45778 + 0.07787X_1 - 0.09547X_2$

The following is the standard error for the intercept: $s(\beta_0) = 2.91534$

The following is the standard error for the Age: $s(\beta_1) = 0.02970$

The following is the standard error for the Health Awareness: $s(\beta_2) = 0.03241$

2 c)

```
pred_55 = data.frame(t(c(55, 60)))
colnames(pred_55) = c('Age', 'Health Awareness')
probit=function(x){
  return(exp(x)/(1+exp(x)))
}
predict(step, pred_55, type='response')
```

```
##          1
## 0.05199847
```

The probability that a 55 year old with Health Awareness index 60 will get a flu shot is 0.05199847.

```
result = predict(step, pred_55, se.fit = TRUE)
c(probit(result$fit-1.96*result$se.fit), probit(result$fit+1.96*result$se.fit))
```

```
##          1          1
## 0.02234752 0.11631025
```

A randomly chosen 55 year old with Health Awareness index 60 will get a flu shot with confidence interval (0.02234752, 0.11631025) at $\alpha = 0.05$.

The confidence interval from part 1c is (0.02470225, 0.15680208). The two confidence intervals are quite similar, they both have rather close lower and upper bounds. This makes sense, since it was previously shown that gender is not a good predictor in the previous $\beta_3 = 0$ hypothesis test.

3 a)

```
library(readxl)
geriatrics <- read_excel('GeriatricStudy.xlsx')
geriatrics
```

```
## # A tibble: 100 x 5
##         Y    X1    X2    X3    X4
##    <dbl> <dbl> <dbl> <dbl> <dbl>
##  1    1.    1.    0.   45.   70.
##  2    1.    1.    0.   62.   66.
##  3    2.    1.    1.   43.   64.
##  4    0.    1.    1.   76.   48.
##  5    2.    1.    0.   51.   72.
##  6    1.    1.    1.   73.   39.
##  7    0.    1.    1.   40.   54.
##  8    0.    1.    0.   66.   37.
##  9    2.    1.    1.   80.   81.
## 10    2.    1.    1.   56.   60.
## # ... with 90 more rows
```

```
fit.poisson <- glm(Y ~ ., data = geriatrics, family = poisson())
summary(fit.poisson)
```

```
## 
## Call:
## glm(formula = Y ~ ., family = poisson(), data = geriatrics)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.1854  -0.7819  -0.2564   0.5449   2.3626  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  0.489467   0.336869   1.453  0.14623    
## X1          -1.069403   0.133154  -8.031 9.64e-16 ***
## X2          -0.046606   0.119970  -0.388  0.69766    
## X3           0.009470   0.002953   3.207  0.00134 ** 
## X4           0.008566   0.004312   1.986  0.04698 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 199.19  on 99  degrees of freedom
## Residual deviance: 108.79  on 95  degrees of freedom
## AIC: 377.29
## 
## Number of Fisher Scoring iterations: 5
```

The estimate for β_0 :

```
fit.poisson$coefficients[1]
```

```
## (Intercept) 
##   0.4894672
```

The estimate for β_1 :

```
fit.poisson$coefficients[2]
```

```
##        X1 
## -1.069403
```

The estimate for β_2 :

```
fit.poisson$coefficients[3]
```

```
##          X2
## -0.04660606
```

The estimate for β_3 :

```
fit.poisson$coefficients[4]
```

```
##          X3
## 0.009469987
```

The estimate for β_4 :

```
fit.poisson$coefficients[5]
```

```
##          X4
## 0.008565829
```

Below are the standard errors for each of the coefficients:

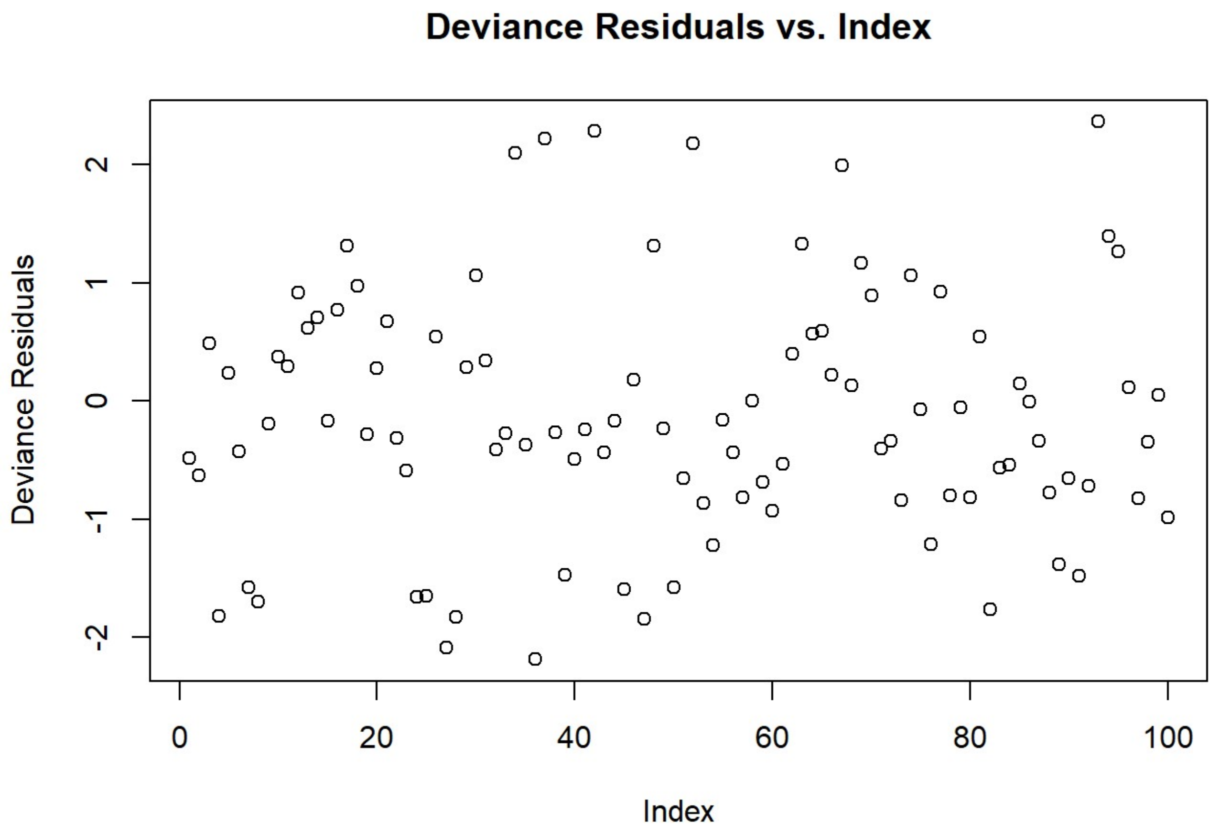
$s(\beta_0) = 0.336869$, $s(\beta_1) = 0.133154$, $s(\beta_2) = 0.119970$, $s(\beta_3) = 0.002953$, $s(\beta_4) = 0.004312$

Below is the estimated regression function:

$$\mu = exp(0.489467 - 1.069403X_1 - 0.046606X_2 + 0.009470X_3 + 0.008566X_4)$$

3 b)

```
res.D=residuals(fit.poisson, type='deviance')
plot(res.D, main = 'Deviance Residuals vs. Index', xlab = 'Index', ylab = 'Deviance Residuals')
```



There don’t appear to be any serious outliers. All the deviance residuals are within (-2.5, 2.5).

3 c) $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$

```
fit.poisson2 <- glm(Y ~ X1+X3+X4, data = geriatrics, family = poisson())
fit.poisson
```

```
##
## Call:  glm(formula = Y ~ ., family = poisson(), data = geriatrics)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##    0.489467    -1.069403    -0.046606     0.009470     0.008566
##
## Degrees of Freedom: 99 Total (i.e. Null);  95 Residual
## Null Deviance:       199.2
## Residual Deviance: 108.8     AIC: 377.3
```

```
fit.poisson2
```

```
##
## Call:  glm(formula = Y ~ X1 + X3 + X4, family = poisson(), data = geriatrics)
##
## Coefficients:
## (Intercept)           X1           X3           X4
##    0.443890    -1.077770     0.009471     0.008979
##
## Degrees of Freedom: 99 Total (i.e. Null);  96 Residual
## Null Deviance:       199.2
## Residual Deviance: 108.9     AIC: 375.4
```

```
test_stat = 108.9 - 108.8
test_stat>qchisq(0.95,1)
```

```
## [1] FALSE
```

The p-value is below:

```
1-pchisq(test_stat,1)
```

```
## [1] 0.7518296
```

So we fail to reject the null hypothesis that $\beta_2 = 0$ at $\alpha = 0.05$.

3 d)

```
summary(fit.poisson2)
```

```
##
## Call:
## glm(formula = Y ~ X1 + X3 + X4, family = poisson(), data = geriatrics)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2152  -0.7512  -0.2594   0.5830   2.2893
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.443890   0.317289   1.399  0.16181
## X1          -1.077770   0.131415  -8.201 2.38e-16 ***
## X3           0.009471   0.002957   3.203  0.00136 **
## X4           0.008979   0.004190   2.143  0.03209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 199.19  on 99  degrees of freedom
## Residual deviance: 108.94  on 96  degrees of freedom
## AIC: 375.44
##
## Number of Fisher Scoring iterations: 5
```

Below is a 95% confidence interval for β_1 :

```
c(-1.077770-1.96*0.317289,-1.077770+1.96*0.317289)
```

```
## [1] -1.6996564 -0.4558836
```

Since the confidence interval does not include 0, and that the sign is negative. This indicates that the coefficient β_1 does have the ability to reduce the frequency of falls when controlling for balance and strength.