BTRY 6020 Homework V Solution

NAME: student name

NETID: student NetID

DUE DATE: 8:40 am Friday March 31

Question 1.

Health officials wonder why some people get the flu shot while others don't. In a study designed to shed some light on this, researchers asked a random sample of patients if they had gotten aflu shot, recorded their age and gender, and also gave each a written questionnaire designed to evaluate their health awareness index. Data appear in Hwk5Q1DatSp17. Note here that Y = 1 means they received the flu shot and that males were coded as $X_3 = 1$, females coded as $X_3 = 0$.

A) Obtain the maximum likelihood estimators of $\beta_0, \beta_1, \beta_2$ and β_3. State the fitted regression function.

```r
library(readxl)
Q1DF <- read_excel("Hwk5Q1DatSp17.xlsx")
Q1_logit <- glm(FShot ~ Age + Ind + Gen, data = Q1DF, family = "binomial")
summary(Q1_logit)
```

```
##
## Call:
## glm(formula = FShot ~ Age + Ind + Gen, family = "binomial", data = Q1DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4037  -0.5637  -0.3352  -0.1542   2.9394
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## Age          0.07279    0.03038   2.396  0.01658 *
## Ind         -0.09899    0.03348  -2.957  0.00311 **
## Gen          0.43397    0.52179   0.832  0.40558
## ---
## 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: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

The MLE of regression coefficients are $\hat{\beta}_0 = 1.17716, \hat{\beta}_1 = 0.07279, \hat{\beta}_2 = 0.09899$ and $\hat{\beta}_3 = 0.43397$.

B) What is the estimated probability of getting the flu shot that a male clients aged 55 years with a health awareness score of 60?

```
predict(Q1_logit, data.frame(Age = 55, Ind = 60, Gen = 1), type = "response")
```

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

The estimated probability of getting the flu shot is 0.06422197.

C) Obtain the VIFs for the regression predictors. What conclusions can you reach from these statistics?

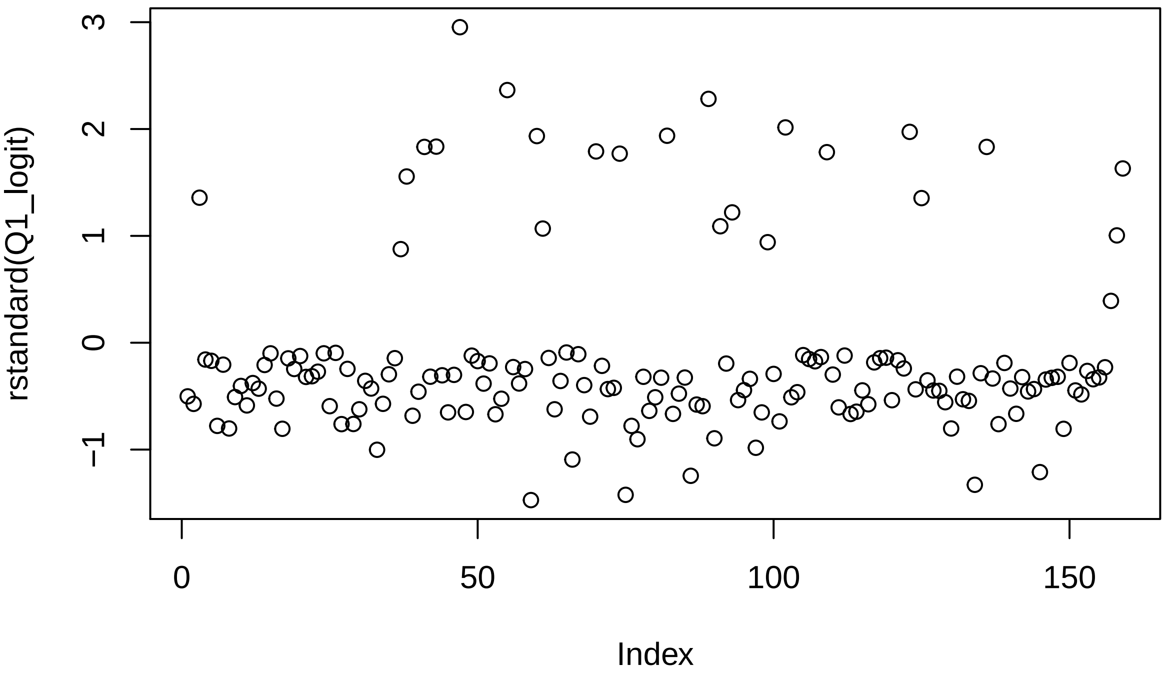
```
library(car)
vif(Q1_logit)
```

```
##      Age      Ind      Gen
## 1.091111 1.081049 1.048432
```

All VIFs are within 2 and therefore multicollinearity is not an issue.

D) Get the standardized deviance residuals and plot against observation number. Does there appear to be any outliers?

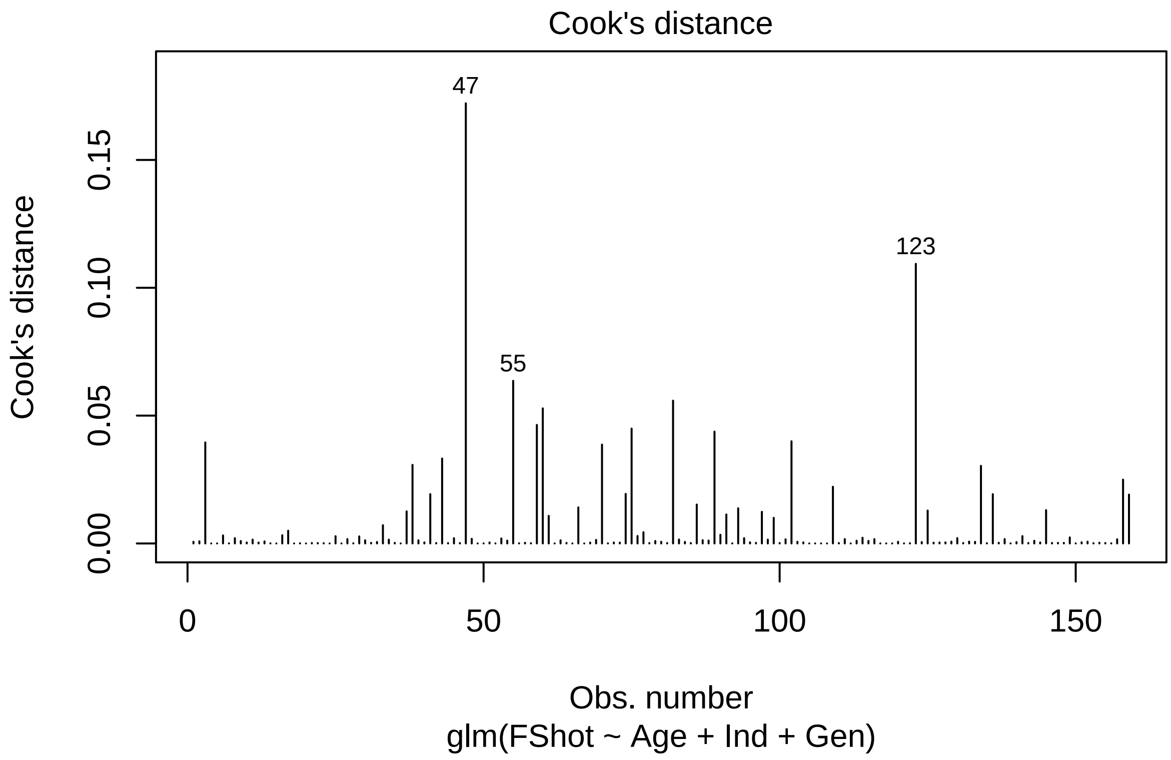
```
plot(rstandard(Q1_logit))
```



The 47th observation has large standardized residual, and sticks out above the rest, which may indicate it's an influential outlier.

E) Get the Cook's distance numbers and plot against observation number. Do there appear to be any influential outliers? If so, check their effects.

```
plot(Q1_logit, which = 4)
```



The 47th and the 123rd observations have large Cook's distance values, making them influential outliers. To evaluate the effects of these two outliers, we refit the logitstic regression after removing them from the data.

```
Q1_logit2 <- glm(FShot ~ Age + Ind + Gen, data = Q1DF[-c(47, 123), ], family = "binomial")
summary(Q1_logit2)
```

```
##
## Call:
## glm(formula = FShot ~ Age + Ind + Gen, family = "binomial", data = Q1DF[-c(47,
##     123), ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5012  -0.5064  -0.2867  -0.1147   2.5339
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93040    3.33632  -0.579  0.56286
## Age          0.08918    0.03379   2.639  0.00831 **
## Ind         -0.11168    0.03771  -2.962  0.00306 **
## Gen          0.75579    0.57566   1.313  0.18922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.23  on 156  degrees of freedom
## Residual deviance:  91.81  on 153  degrees of freedom
## AIC: 99.81
##
```

```
## Number of Fisher Scoring iterations: 6
```

After removing the two outliers, the p-values of all three predictors become smaller which establishes more significance in these variables, though gender remains nonsignificant. The regression coefficients also got reinforced in respective direction, with the coefficient of Age increasing by 22.5% and that of index decreasing (becoming more significantly negative) by 14.1%.

F) Can we drop Age and Gender if we keep the health awareness index in the model? State hypotheses, test statistic, p-value, and conclusions.

The null and alternative hypotheses are as follows:

$$H_0 : \beta_1 = \beta_3 = 0 \quad \text{and} \quad H_a : \text{At least one of } \beta_1 \text{ and } \beta_3 \text{ is not zero}$$

To test the hypotheses, we may conduct a likelihood ratio test where the full model is listed in part A) and the reduced model is

```
Q1_logit_reduced <- glm(FShot ~ Ind, data = Q1DF, family = "binomial")
anova(Q1_logit_reduced, Q1_logit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: FShot ~ Ind
## Model 2: FShot ~ Age + Ind + Gen
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       157     113.20
## 2       155     105.09  2   8.1026   0.0174 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With p-value being 0.0174 less than 0.05, we reject H_0 in favor of H_a that at least one of Age and Gender is significant in explaining variations in the response. Hence, we cannot drop Age and Gender while keeping the health awareness index in the model.

G) Install the package “bestglm”. Visit the following website:

https://cran.r-project.org/web/packages/bestglm/vignettes/bestglm.pdf

to learn how to use this package. Don’t forget the “library(bestglm)” command before you use it.

```
i) Find the best model for getting a flu shot according to the BIC criteria
```

```
library(bestglm)
bestBIC <- bestglm(Q1DF[c("Age", "Ind", "Gen", "FShot")], family = binomial, IC = "BIC")
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
bestBIC
```

```
## BIC
## BICq equivalent for q in (0.237623513993524, 0.898744026033073)
## Best Model:
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -1.45778309 2.91533637 -0.5000394 0.617047325
## Age          0.07787235 0.02969670  2.6222563 0.008734971
## Ind         -0.09547230 0.03240764 -2.9459813 0.003219318
```

The best model using BIC criteria is

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Age + \beta_2 Ind$$

ii) Find the best models for a 0, 1, 2, and 3 predictors using the Subsets command

```
bestBIC$Subsets
```

```
##    Intercept   Age   Ind   Gen logLikelihood      BIC
## 0       TRUE FALSE FALSE FALSE     -67.47038 134.9408
## 1       TRUE FALSE  TRUE FALSE     -56.59790 118.2647
## 2*      TRUE  TRUE  TRUE FALSE     -52.89769 115.9332
## 3       TRUE  TRUE  TRUE  TRUE     -52.54659 120.2999
```

```
# 0 predictor
bestglm(Q1DF[c("Age", "Ind", "Gen", "FShot")], family = binomial, IC = "BIC", nvmax = 0)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## BIC
## Best Model:
##              Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -1.727221  0.2215267 -7.796898 6.344756e-15
```

```
# 1 predictor
bestglm(Q1DF[c("Age", "Ind", "Gen", "FShot")], family = binomial, IC = "BIC", nvmax = 1)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## BIC
## Best Model:
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  4.9113285 1.62651185  3.019547 2.531533e-03
## Ind         -0.1193093 0.03012988 -3.959832 7.500247e-05
```

```
# 2 predictors
bestglm(Q1DF[c("Age", "Ind", "Gen", "FShot")], family = binomial, IC = "BIC", nvmax = 2)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## BIC
## Best Model:
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -1.45778309 2.91533637 -0.5000394 0.617047325
## Age          0.07787235 0.02969670  2.6222563 0.008734971
## Ind         -0.09547230 0.03240764 -2.9459813 0.003219318
```

```
# 3 predictors
summary(glm(FShot ~ Age + Ind + Gen, data = Q1DF, family = "binomial"))
```

```
##
## Call:
## glm(formula = FShot ~ Age + Ind + Gen, family = "binomial", data = Q1DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4037  -0.5637  -0.3352  -0.1542   2.9394
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## Age          0.07279    0.03038   2.396  0.01658 *
## Ind         -0.09899    0.03348  -2.957  0.00311 **
## Gen          0.43397    0.52179   0.832  0.40558
## ---
## 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: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

iii) Find the best model for getting a flu shot according to the AIC criteria

```
bestAIC <- bestglm(Q1DF[c("Age", "Ind", "Gen", "FShot")], family = binomial, IC = "AIC")
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
bestAIC
```

```
## AIC
## BICq equivalent for q in (0.237623513993524, 0.898744026033073)
## Best Model:
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -1.45778309 2.91533637 -0.5000394 0.617047325
## Age          0.07787235 0.02969670  2.6222563 0.008734971
## Ind         -0.09547230 0.03240764 -2.9459813 0.003219318
```

The best model using AIC criteria is

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Age + \beta_2 Ind$$

iv) Find the best models for a 0, 1, 2, and 3 predictors using the Subsets command

```
bestAIC$Subsets
```

```
##    Intercept   Age   Ind   Gen logLikelihood      AIC
## 0       TRUE FALSE FALSE FALSE     -67.47038 134.9408
## 1       TRUE FALSE  TRUE FALSE     -56.59790 115.1958
## 2*      TRUE  TRUE  TRUE FALSE     -52.89769 109.7954
## 3       TRUE  TRUE  TRUE  TRUE     -52.54659 111.0932
```

```
# 0 predictor
bestglm(Q1DF[c("Age", "Ind", "Gen", "FShot")], family = binomial, IC = "AIC", nvmax = 0)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## AIC
## Best Model:
##              Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -1.727221  0.2215267 -7.796898 6.344756e-15
```

```
# 1 predictor
bestglm(Q1DF[c("Age", "Ind", "Gen", "FShot")], family = binomial, IC = "AIC", nvmax = 1)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## AIC
## Best Model:
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  4.9113285 1.62651185  3.019547 2.531533e-03
## Ind         -0.1193093 0.03012988 -3.959832 7.500247e-05
```

```r
# 2 predictors
bestglm(Q1DF[c("Age", "Ind", "Gen", "FShot")], family = binomial, IC = "AIC", nvmax = 2)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## AIC
## Best Model:
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -1.45778309 2.91533637 -0.5000394 0.617047325
## Age          0.07787235 0.02969670  2.6222563 0.008734971
## Ind         -0.09547230 0.03240764 -2.9459813 0.003219318
```

```r
# 3 predictors
summary(glm(FShot ~ Age + Ind + Gen, data = Q1DF, family = "binomial"))
```

```
##
## Call:
## glm(formula = FShot ~ Age + Ind + Gen, family = "binomial", data = Q1DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4037  -0.5637  -0.3352  -0.1542   2.9394
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## Age          0.07279    0.03038   2.396  0.01658 *
## Ind         -0.09899    0.03348  -2.957  0.00311 **
## Gen          0.43397    0.52179   0.832  0.40558
## ---
## 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: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

v) What model from the above models evaluated would you choose for this situation? Explain BRIEFLY; you may include data from all parts of Question 1.

I will choose $log(\frac{p}{1-p}) = \beta_0 + \beta_1 Age + \beta_2 Ind$ since it's the best model using both AIC and BIC criteria.

Question 2.

A disease outbreak has occurred in a certain city. Data have been collected on a random telephone survey of 196 people within city limits and the following data recorded: 1) Whether or not they have contracted the disease (Dis, =1 if they have, =0 if not), Age, Socioeconomic Status (SES, = 1 if upper, = 2 if middle, = 3 if lower), Sector of the city they live (Sect, either sector 1 or sector 2), and saving account status (Sav, = 1 if they have a savings account, = 0 if not). data appear in Hwk5Q2DatSp17

Part A) Develop a logistic regression model for predicting the probability of contracting this disease, using the above variables. Be sure to check for polynomial effects of significant quantitative variables as well as interactions between significant predictor variables. When finished, explicitly state your prediction equation. Be sure to show significant steps in model development, using simultaneous tests when you want to omit/test more than one predictor.

First, we consider up to cubic polynomial terms of Age, all three categorical variables (SES, Sect and Sav) and the interactions between Age and the categorical variables. The following is the logistic fit.

```r
# Read the data
Q2DF <- read_excel("Hwk5Q2DatSp17.xlsx")
Q2DF$SES <- factor(Q2DF$SES, levels = c(1,2,3))
Q2DF$Sect <- factor(Q2DF$Sect, levels = c(1, 2))
Q2DF$Sav <- factor(Q2DF$Sav, levels = c(0, 1))
Q2_logit <- glm(Dis ~ Age + I(Age^2) + I(Age^3) + Age:SES + SES + Age:Sect + Sect + Age:Sav + Sav, data
summary(Q2_logit)
```

```
##
## Call:
## glm(formula = Dis ~ Age + I(Age^2) + I(Age^3) + Age:SES + SES +
##     Age:Sect + Sect + Age:Sav + Sav, family = binomial, data = Q2DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5567  -0.7827  -0.4906   0.9344   2.2266
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.290e+00  1.294e+00  -3.316 0.000914 ***
## Age          2.170e-01  9.131e-02   2.376 0.017499 *
## I(Age^2)    -4.788e-03  2.572e-03  -1.862 0.062663 .
## I(Age^3)     3.060e-05  2.141e-05   1.430 0.152810
## SES2        -4.203e-01  9.848e-01  -0.427 0.669564
## SES3         6.248e-01  8.779e-01   0.712 0.476618
## Sect2        9.755e-01  7.491e-01   1.302 0.192837
## Sav1         3.418e-01  8.789e-01   0.389 0.697344
## Age:SES2     1.874e-02  2.911e-02   0.644 0.519616
## Age:SES3     4.030e-04  2.518e-02   0.016 0.987231
## Age:Sect2    1.468e-02  2.306e-02   0.637 0.524348
## Age:Sav1    -3.409e-03  3.065e-02  -0.111 0.911433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 199.18  on 184  degrees of freedom
## AIC: 223.18
##
## Number of Fisher Scoring iterations: 5
```

From the above fit, we consider removing all the interaction terms by conducting a LRT. The hypotheses and LRT test are as follows:

$$H_0 : \text{the partial slopes associated with all the interactions are all zero.}$$

$$H_a : \text{at least one of the partial slopes associated with the interactiosn is not zero.}$$

```
Q2_logit2 <- glm(Dis ~ Age + I(Age^2) + I(Age^3) + SES + Sect + Sav, data = Q2DF, family = binomial)
summary(Q2_logit2)
```

```
##
## Call:
## glm(formula = Dis ~ Age + I(Age^2) + I(Age^3) + SES + Sect +
##     Sav, family = binomial, data = Q2DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5419  -0.8053  -0.4894   0.9786   2.3486
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.568e+00  1.040e+00  -4.393 1.12e-05 ***
## Age          2.208e-01  8.630e-02   2.559 0.010503 *
## I(Age^2)    -4.493e-03  2.490e-03  -1.804 0.071200 .
## I(Age^3)     2.735e-05  2.062e-05   1.326 0.184720
## SES2         1.131e-01  4.564e-01   0.248 0.804344
## SES3         6.166e-01  4.699e-01   1.312 0.189487
## Sect2        1.428e+00  3.772e-01   3.787 0.000152 ***
## Sav1         2.320e-01  4.245e-01   0.547 0.584718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 200.36  on 188  degrees of freedom
## AIC: 216.36
##
## Number of Fisher Scoring iterations: 5
```

```
anova(Q2_logit2, Q2_logit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Dis ~ Age + I(Age^2) + I(Age^3) + SES + Sect + Sav
## Model 2: Dis ~ Age + I(Age^2) + I(Age^3) + Age:SES + SES + Age:Sect +
##     Sect + Age:Sav + Sav
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       188     200.36
## 2       184     199.18  4   1.1853   0.8805
```

The p-value associated with the LRT is 0.8805 and hence we fail to reject H_0. Therefore, we can drop all the interaction terms. From the summary tabel for Q2_logit2, we notice Age^3, SES and Sav are not significant. We will do another LRT to test if we can drop them all at once.

$$H_0 : \text{the partial slopes associated with } Age^3, SES \text{ and } Sav \text{ are all zero.}$$
$$H_a : \text{at least one of the partial slopes associated with } Age^3, SES \text{ and } Sav \text{ is not zero.}$$

```r
Q2_logit3 <- glm(Dis ~ Age + I(Age^2) + Sect, data = Q2DF, family = binomial)
summary(Q2_logit3)
```

```
## 
## Call:
## glm(formula = Dis ~ Age + I(Age^2) + Sect, family = binomial, 
##     data = Q2DF)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.4809  -0.8260  -0.5211   0.9359   2.2128  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -3.281981   0.582123  -5.638 1.72e-08 ***
## Age          0.113425   0.033833   3.352 0.000801 ***
## I(Age^2)    -0.001194   0.000446  -2.677 0.007427 ** 
## Sect2        1.278602   0.348733   3.666 0.000246 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 203.58  on 192  degrees of freedom
## AIC: 211.58
## 
## Number of Fisher Scoring iterations: 4
```

```r
anova(Q2_logit3, Q2_logit2, test = "LRT")
```

```
## Analysis of Deviance Table
## 
## Model 1: Dis ~ Age + I(Age^2) + Sect
## Model 2: Dis ~ Age + I(Age^2) + I(Age^3) + SES + Sect + Sav
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       192     203.57                     
## 2       188     200.36  4    3.212    0.523
```

With p-value being 0.523, we fail to reject H_0 and drop Age^3, SES and Sav all at once.The final model equation is

$$\log\left(\frac{\hat{p}_{disease}}{1-\hat{p}_{disease}}\right) = -3.282 + 0.113Age - 0.001Age^2 + 1.279Sect$$

Part B) Give a 90% confidence interval for the probability that a 64 year old patient, with middle socioeconomic status and a savings account that lives in sector 2 of the city, contracts the disease.

We use our final model to predict the porbability of that person contracting the disease:

```r
predict(Q2_logit3, newdata = data.frame(Age = 64, Sect = as.factor(2)), se.fit = T, type = "link")
```

```
## $fit
##       1
## 0.36516
##
## $se.fit
## [1] 0.4239494
##
## $residual.scale
## [1] 1
```

A 90% confidence interval for the log-odds is $(0.365 - 1.645 * 0.4239, 0.365 + 1.645 * 0.4239) = (-0.3323, 1.0623)$. Thus, a 90% confidence interval for $p_{disease}$ is $\left(\frac{e^{-0.3323}}{1+e^{-0.3323}}, \frac{e^{1.0623}}{1+e^{1.0623}}\right) = (0.41768, 0.74313)$.

Question 3.

Multiple cohorts of subjects, some non-smokers and others smokers, were observed for several years. The number of cases (NumCases) of lung cancer diagnosed in the different cohorts was recorded, in addition to the following predictor variables:

CigsperDay = Number of cigarettes smoked per day per individual in the cohort; Years = The number of years the individuals in the cohort had smoked.

Additionally, the total number of years in which individuals in each category were observed (summed over all individuals) was recorded in the column PersonYears. (For example, if a cohort had 50 people that had been observed for 20 years, that would be 50 x 20 = 1000 PersonYears.) Data appear in Hwk5Q3DatSp17.

A) Write down a Poisson regression model where the mean number of cases of observed lung cancer cases per cohort are a function of CigsperDay and Years. Your model should start like "μ? = .", NOT "log(μ) = .".

$$E[NumCases] = \exp\left(\beta_0 + \beta_1 \times CigsperDay + \beta_2 \times Years\right).$$

B) Fit the model above; include summary output. State your model of the estimated mean with the maximum likelihood estimators included.

```r
library(readxl)
Q3DF <- read_excel("Hwk5Q3DatSp17.xlsx")
Q3_poisson <- glm(NumCases ~ CigsperDay + Years, data = Q3DF, family = poisson)
summary(Q3_poisson)
```

```
##
## Call:
## glm(formula = NumCases ~ CigsperDay + Years, family = poisson,
##     data = Q3DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6501  -1.4275  -0.9166   0.4112   3.2095
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.726048   0.450482  -3.832 0.000127 ***
## CigsperDay   0.040434   0.008561   4.723 2.32e-06 ***
## Years        0.043769   0.008942   4.895 9.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 137.29  on 34  degrees of freedom
## Residual deviance:  88.13  on 32  degrees of freedom
## AIC: 151.72
##
## Number of Fisher Scoring iterations: 6
```

The fitted model is

$$\hat{E}[NumCases] = \exp\left(-1.726048 + 0.040434 CigsperDay + 0.043769 Years\right).$$

C) Do a deviance goodness-of-fit test on your model; state hypotheses, test statistic, p-value, and conclusions.

$$H_0 : \text{the data are consistent under the specified model}$$
$$H_a : \text{the data is not consistent under the specified model.}.$$

The test statistic is residual deviance 88.13, which under the null follows a chisquare distribution with 32 degrees of freedom. The p-value is computed below.

```
1-pchisq(88.13, 32)
```

```
## [1] 3.814793e-07
```

Since p-value is less than 0.05, we reject the null hypothesis and conclude that our model dosn't fit the data well.

D) Does it make sense for your mean in Part A above to be proportional to the variable PersonYears? Explain briefly.

Yes, the larger PersonYears in a cohort is, the more participants there are in a cohort. In a larger cohort we are more likely to observe more lung cancer cases, provided everything else stay the same. Therefore, the mean in Part A should be proportional to PersonYears.

E) Write down a Poisson regression model where the mean number of cases of observed lung cancer cases per cohort are a function of CigsperDay and Years, but are also proportional to PersonYears. Your model should start like "? = .", NOT "log(?) = ."

$$E[NumCases] = \exp\left(\beta_0 + \beta_1 \times CigsperDay + \beta_2 \times Years\right) \times PersonYears.$$

F) Fit the above model; include summary output. Perform a deviance goodness-of-fit test on this model; state hypotheses, test statistic, p-value, and conclusions.

Note here we set offset as $0.001 \times PersonYears$ in order for the glm algorithm to converge. This change won't invalidate the previous setup that the mean number of cases of observed lung cancer cases per cohort is proportional to PersonYears.

```
Q3_poisson2 <- glm(NumCases ~ CigsperDay + Years, data = Q3DF, family = poisson, offset = log(PersonYea
summary(Q3_poisson2)
```

```
##
## Call:
## glm(formula = NumCases ~ CigsperDay + Years, family = poisson,
##     data = Q3DF, offset = log(PersonYears))
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1657  -1.1254  -0.5335   0.5965   1.4920
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.557675   0.546276 -22.988  < 2e-16 ***
## CigsperDay    0.070795   0.009415   7.519 5.51e-14 ***
## Years         0.120894   0.010760  11.235  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.712  on 34  degrees of freedom
## Residual deviance:  43.347  on 32  degrees of freedom
## AIC: 106.94
##
## Number of Fisher Scoring iterations: 5
```

$$H_0 : \text{the data are consistent under the specified model}$$

$$H_a : \text{the data is not consistent under the specified model..}$$

The test statistic is residual deviance 88.13, which under the null follows a chisquare distribution with 32 degrees of freedom. The p-value is computed below.

```
1-pchisq(106.48, 32)
```

```
## [1] 6.234572e-10
```

Since p-value is less than 0.05, we reject the null hypothesis and conclude that our model dosn't fit the data well. In fact, adding the reasonable offset seems to have made the model worse, with the residual deviance rising from 88.13 to 137.29. But you must consider the null deviance with these two different models. The R^2 analog in generalized linear models is 1-(Residual Deviance/Null Deviance). Without the offset, this equals 1-(88.13/137.29) = 35.8%, a small amount of the deviance in the number of cases. With the offset, this R^2 rises to 1-(43.347/250.712)=82.7% of the deviance in the number of cases, a MUCH better model.

However, the lack of fit is still significant, and at this point if we're to do inferences, we should go back to a quasipoisson model and start over.