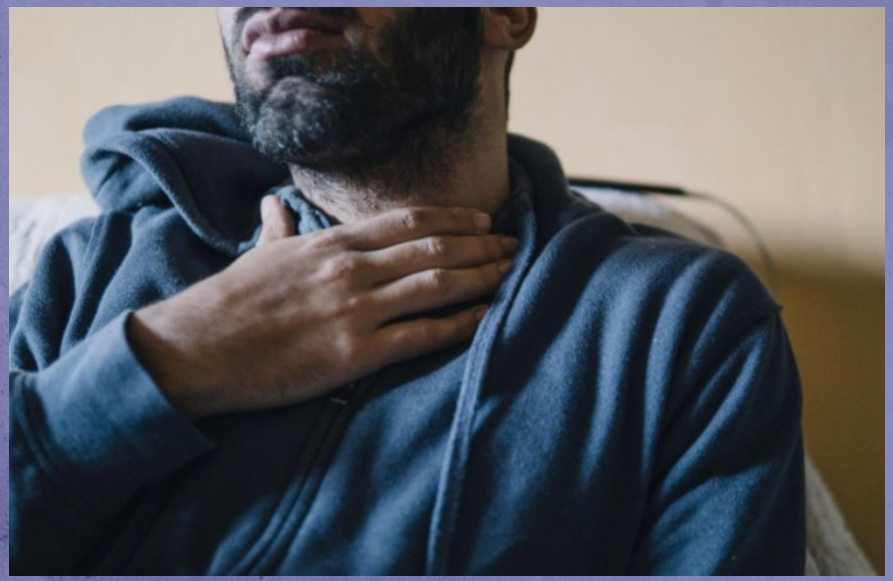
ESOPHAGEAL CANCER STUDY

EFE UMUKORO

LOGISTIC REGRESSION PROJECT

BACKGROUND



Esophageal Cancer is a cancer that occurs in the esophagus. The esophagus is a long hollow tub that runs from your throat to your stomach aiding in the movement of food once it is swallowed.

One of the 6th common causes of morbidity in the United States is due to Esophageal cancer. The incidence of morbidity is most common among men than in women

Common risk factors include use of tobacco, alcohol consumption, nutritional habits and obesity. The goal of this project is to investigate alcohol, diet, tobacco consumption as potential risk factors for esophageal cancer in men.

DATA

The data contains the following variables:

Case: 1=case, 0 = control

Age: age in years

Alc: alcohol consumption amount

Alcgp: alchohol group categorization

Tob: tobacco consumption

Tobgp: smoking categorization

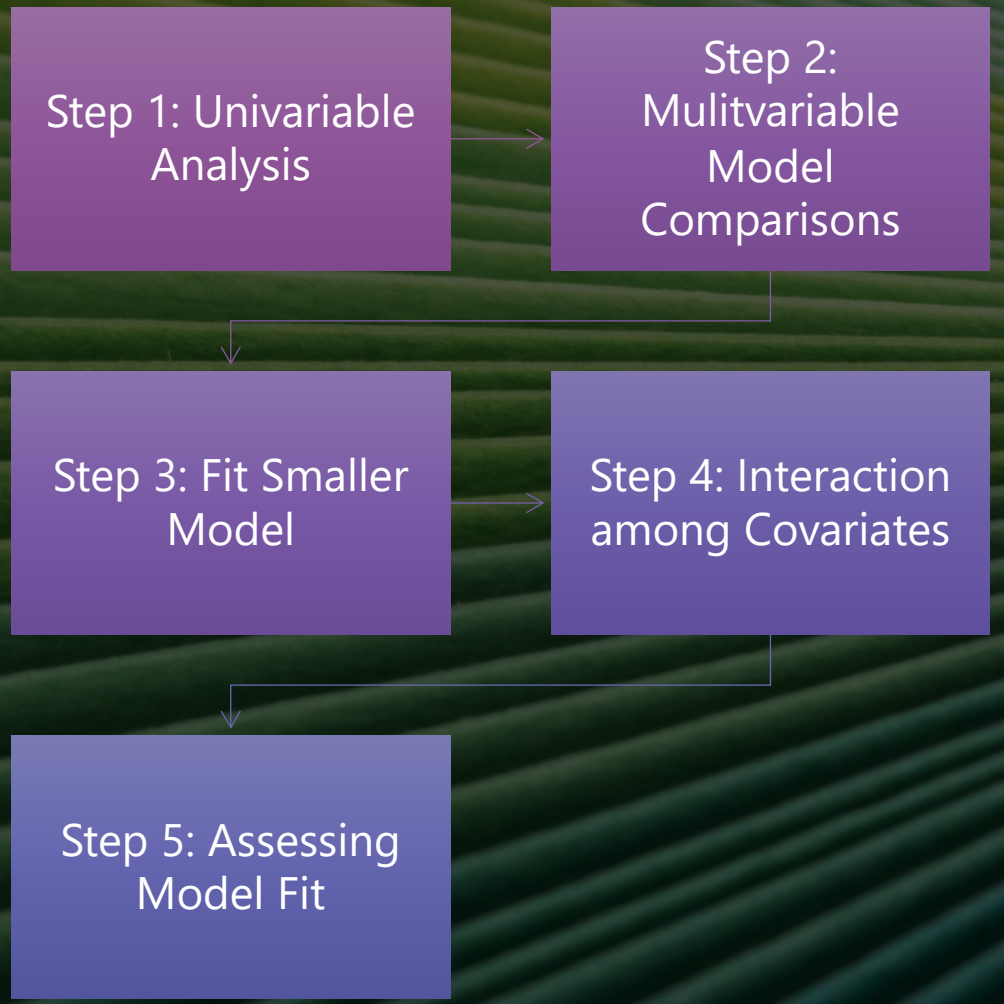
RESEARCH QUESTION

Given the variables and data available, What are the risk factors that influence the presence or absence of esophageal cancer ?

STATISTICAL METHOD

Logistic Regression

LOGISTIC REGRESSION MODEL CREATION

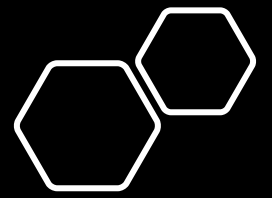


	case	age	agegp	tob	tobgp	alc	alcgp
	<dbl>	<dbl>	<dbl+lbl>	<dbl>	<dbl+lbl>	<dbl>	<dbl+lbl>
1	0	42	2 [35-44]	0	1 [0-9 gm/day]	139	4 [120+]
2	0	45	3 [45-54]	7.5	1 [0-9 gm/day]	66	2 [40-79]
3	0	35	2 [35-44]	0	1 [0-9 gm/day]	24	1 [0-39 gm/day]
4	0	78	6 [75+]	0	1 [0-9 gm/day]	39	1 [0-39 gm/day]
5	0	45	3 [45-54]	7.5	1 [0-9 gm/day]	64	2 [40-79]
6	0	64	4 [55-64]	17.5	2 [10-19]	49	2 [40-79]
7	0	76	6 [75+]	2.5	1 [0-9 gm/day]	1	1 [0-39 gm/day]
8	0	42	2 [35-44]	0	1 [0-9 gm/day]	33	1 [0-39 gm/day]
9	0	48	3 [45-54]	12.5	2 [10-19]	55	2 [40-79]
10	0	42	2 [35-44]	25	3 [20-29]	62	2 [40-79]

… with 965 more rows

```
> summary(iv)
      case              age             agegp            tob              tobgp      
 Min.   :0.0000   Min.   :25.00   Min.   :1.000   Min.   : 0.00   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:41.00   1st Qu.:2.000   1st Qu.: 0.00   1st Qu.:1.000  
 Median :0.0000   Median :52.00   Median :3.000   Median : 7.50   Median :1.000  
 Mean   :0.2051   Mean   :52.23   Mean   :3.272   Mean   :11.75   Mean   :1.763  
 3rd Qu.:0.0000   3rd Qu.:63.00   3rd Qu.:4.000   3rd Qu.:17.50   3rd Qu.:2.000  
 Max.   :1.0000   Max.   :91.00   Max.   :6.000   Max.   :60.00   Max.   :4.000  
      alc              alcgp      
 Min.   :  0.00   Min.   :1.000  
 1st Qu.: 24.00   1st Qu.:1.000  
 Median : 45.00   Median :2.000  
 Mean   : 52.77   Mean   :1.855  
 3rd Qu.: 73.00   3rd Qu.:2.000  
 Max.   :268.00   Max.   :4.000  
```

OVERVIEW OF DATA AND VARIABLES



```
> age.model = glm(case ~ age, family = binomial)
> age.model

Call:  glm(formula = case ~ age, family = binomial)

Coefficients:
(Intercept)          age
   -4.42748      0.05562

Degrees of Freedom: 974 Total (i.e. Null);  973 Residual
Null Deviance:      989.5
Residual Deviance: 906.6        AIC: 910.6
> summary(age.model)

Call:
glm(formula = case ~ age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4559  -0.7431  -0.4955  -0.3412   2.2858

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.427475   0.389388 -11.370   <2e-16 ***
age          0.055624   0.006565   8.473   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 906.64  on 973  degrees of freedom
AIC: 910.64

Number of Fisher Scoring iterations: 4
```

```
> tobacco.model = glm(case ~ tobacco, family = binomial)
> tobacco.model

Call:  glm(formula = case ~ tobacco, family = binomial)

Coefficients:
(Intercept)      tobacco
    -1.8433       0.0367

Degrees of Freedom: 974 Total (i.e. Null);  973 Residual
Null Deviance:      989.5
Residual Deviance: 952.6        AIC: 956.6
> summary(tobacco.model)

Call:
glm(formula = case ~ tobacco, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3331  -0.6686  -0.5657  -0.5421   1.9951

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.843280   0.120023 -15.358  < 2e-16 ***
tobacco      0.036704   0.006015   6.102 1.05e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 952.62  on 973  degrees of freedom
AIC: 956.62

Number of Fisher Scoring iterations: 4
```

```
> alcohol.model = glm(case ~ alcohol, family = binomial)
> alcohol.model

Call:  glm(formula = case ~ alcohol, family = binomial)

Coefficients:
(Intercept)      alcohol
   -2.96890      0.02601

Degrees of Freedom: 974 Total (i.e. Null);  973 Residual
Null Deviance:      989.5
Residual Deviance: 830.4        AIC: 834.4
> summary(alcohol.model)

Call:
glm(formula = case ~ alcohol, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7893  -0.6312  -0.4553  -0.3205   2.4572

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.968901   0.180807  -16.42   <2e-16 ***
alcohol      0.026014   0.002319   11.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 830.41  on 973  degrees of freedom
AIC: 834.41
```

UNIVARIABLE ANALYSIS

INITIAL MULTIVARIATE MODEL MODEL

```
> summary(model.iv)

Call:
glm(formula = case ~ age + tobacco + alcohol, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9323  -0.5923  -0.3206  -0.1410   2.8619

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.572920   0.586349 -12.915  < 2e-16 ***
age          0.072289   0.008222   8.792  < 2e-16 ***
tobacco      0.038803   0.007574   5.123 3.01e-07 ***
alcohol      0.026490   0.002525  10.492  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 718.28  on 971  degrees of freedom
AIC: 726.28

Number of Fisher Scoring iterations: 5
```

```
> model.iv = glm(case ~ age + tobacco + alcohol, family = binomial)
> model.iv

Call:  glm(formula = case ~ age + tobacco + alcohol, family = binomial)

Coefficients:
(Intercept)          age      tobacco      alcohol
   -7.57292      0.07229      0.03880      0.02649

Degrees of Freedom: 974 Total (i.e. Null);  971 Residual
Null Deviance:      989.5
Residual Deviance: 718.3        AIC: 726.3
> summary(model.iv)
```

```
> model.ivint1 = glm(case ~ age + tobacco + alcohol + age:tobacco, family = binomial)
> model.ivint1

Call:  glm(formula = case ~ age + tobacco + alcohol + age:tobacco, family = binomial)

Coefficients:
(Intercept)          age      tobacco      alcohol  age:tobacco
 -6.9155931    0.0607642   -0.0057793    0.0265609    0.0008051

Degrees of Freedom: 974 Total (i.e. Null);  970 Residual
Null Deviance:      989.5
Residual Deviance: 716.4        AIC: 726.4
> summary(model.ivint1)

Call:
glm(formula = case ~ age + tobacco + alcohol + age:tobacco, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9477  -0.5871  -0.3269  -0.1568   2.8277

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.9155931  0.7427999  -9.310  < 2e-16 ***
age          0.0607642  0.0116783   5.203 1.96e-07 ***
tobacco     -0.0057793  0.0344699  -0.168    0.867
alcohol      0.0265609  0.0025313  10.493  < 2e-16 ***
age:tobacco  0.0008051  0.0006066   1.327    0.184
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 716.44  on 970  degrees of freedom
AIC: 726.44

Number of Fisher Scoring iterations: 5
```

```
> model.ivint3 = glm(case ~ age + tobacco + alcohol + tobacco:alcohol, family = binom
> model.ivint3

Call:  glm(formula = case ~ age + tobacco + alcohol + tobacco:alcohol,
    family = binomial)

Coefficients:
    (Intercept)              age          tobacco          alcohol  tobacco:alcohol
     -7.7134345        0.0721749        0.0486391        0.0288391       -0.0001538

Degrees of Freedom: 974 Total (i.e. Null);  970 Residual
Null Deviance:      989.5
Residual Deviance: 717.6        AIC: 727.6
> summary(model.ivint3)

Call:
glm(formula = case ~ age + tobacco + alcohol + tobacco:alcohol,
    family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9768  -0.5949  -0.3244  -0.1371   2.9033

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.7134345  0.6138535 -12.566  < 2e-16 ***
age              0.0721749  0.0082343   8.765  < 2e-16 ***
tobacco          0.0486391  0.0140431   3.464 0.000533 ***
alcohol          0.0288391  0.0038329   7.524 5.31e-14 ***
tobacco:alcohol -0.0001538  0.0001845  -0.834 0.404525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 717.60  on 970  degrees of freedom
AIC: 727.6
```

```
> model.ivint2 = glm(case ~ age + tobacco + alcohol + age:alcohol, family = binomi
> model.ivint2

Call:  glm(formula = case ~ age + tobacco + alcohol + age:alcohol, family = binomi

Coefficients:
(Intercept)          age      tobacco      alcohol  age:alcohol
 -7.0561425    0.0633991    0.0390663    0.0181448    0.0001465

Degrees of Freedom: 974 Total (i.e. Null);  970 Residual
Null Deviance:      989.5
Residual Deviance: 717.7        AIC: 727.7
> summary(model.ivint2)

Call:
glm(formula = case ~ age + tobacco + alcohol + age:alcohol, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0102  -0.5908  -0.3311  -0.1513   2.8275

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.0561425  0.9052709  -7.795 6.47e-15 ***
age          0.0633991  0.0145577   4.355 1.33e-05 ***
tobacco      0.0390663  0.0075710   5.160 2.47e-07 ***
alcohol      0.0181448  0.0116087   1.563    0.118
age:alcohol  0.0001465  0.0001999   0.733    0.463
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 717.74  on 970  degrees of freedom
AIC: 727.74
```

DETERMINE POSSIBLE VARIABLE INTERACTIONS

```
Coefficients:
(Intercept)          age      tobacco      alcohol
   -7.57292      0.07229      0.03880      0.02649

Degrees of Freedom: 974 Total (i.e. Null);  971 Residual
Null Deviance:      989.5
Residual Deviance: 718.3        AIC: 726.3
> summary(iv.finalmodel)

Call:
glm(formula = case ~ age + tobacco + alcohol, family = binomial,
    data = iv)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9323  -0.5923  -0.3206  -0.1410   2.8619

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.572920   0.586349 -12.915  < 2e-16 ***
age          0.072289   0.008222   8.792  < 2e-16 ***
tobacco      0.038803   0.007574   5.123 3.01e-07 ***
```

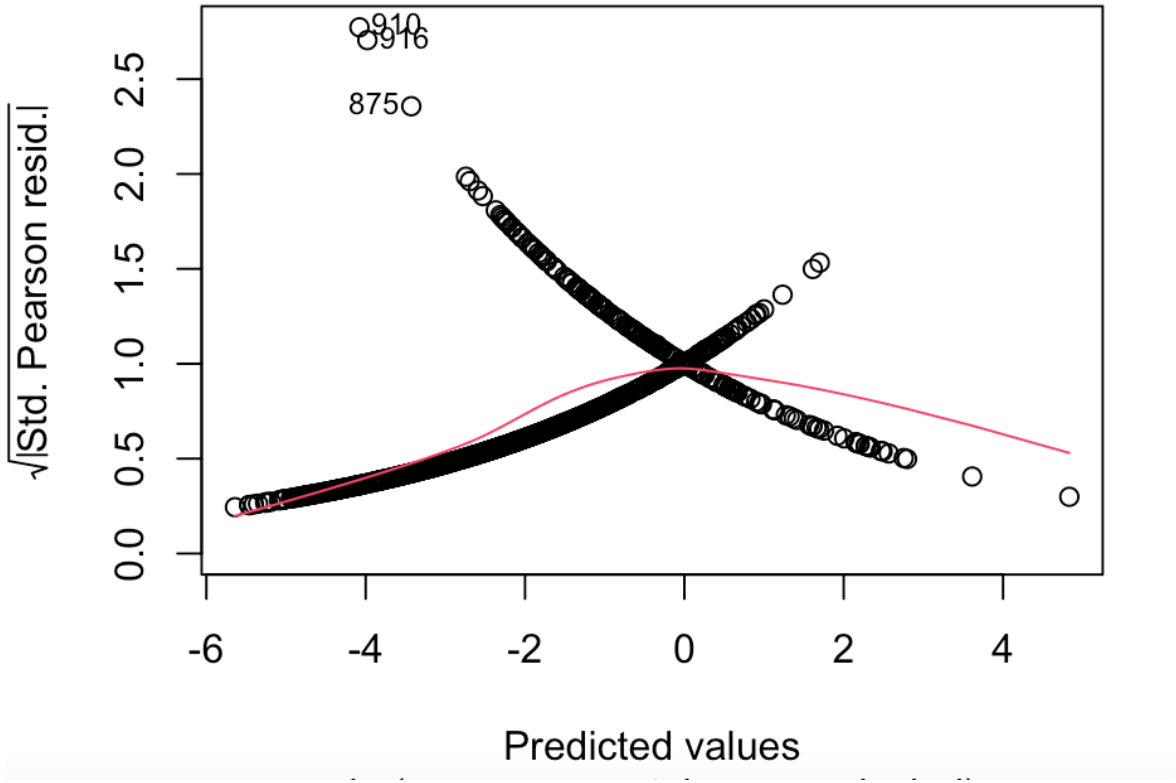
FINAL MODEL

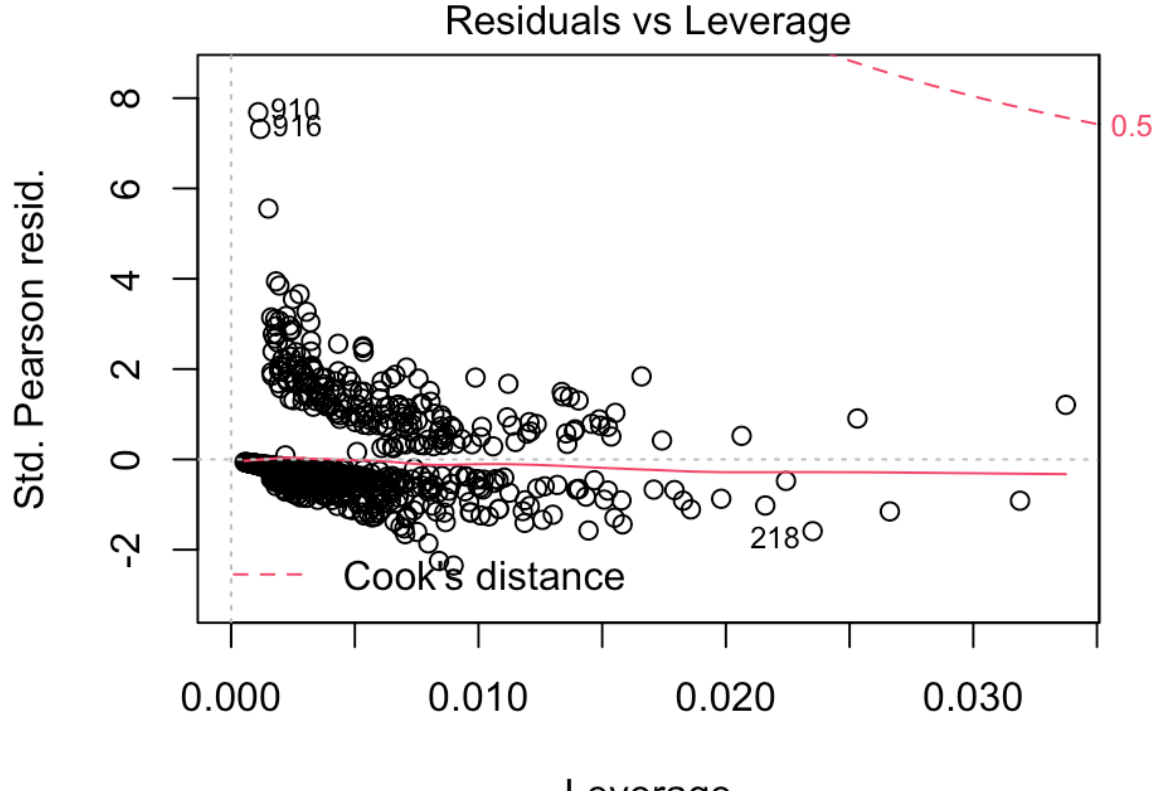
Assessment of Model Fit: Pearson Test

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  ivfinalmodel$y, fitted(ivfinalmodel)
X-squared = 16.195, df = 8, p-value = 0.03967
```

```
> prob1 = 1 - pchisq(718.281, 851.8147)
> prob1
[1] 0.9996695
```





Residual Analysis

```
> ivfinalmodel

Call:  glm(formula = case ~ age + tobacco + alcohol, family = binomial)

Coefficients:
(Intercept)          age      tobacco      alcohol
   -7.57292      0.07229      0.03880      0.02649

Degrees of Freedom: 974 Total (i.e. Null);  971 Residual
Null Deviance:      989.5
Residual Deviance: 718.3        AIC: 726.3
> summary(ivfinalmodel)

Call:
glm(formula = case ~ age + tobacco + alcohol, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9323  -0.5923  -0.3206  -0.1410   2.8619

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.572920   0.586349 -12.915  < 2e-16 ***
age          0.072289   0.008222   8.792  < 2e-16 ***
tobacco      0.038803   0.007574   5.123 3.01e-07 ***
alcohol      0.026490   0.002525  10.492  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 718.28  on 971  degrees of freedom
AIC: 726.28

Number of Fisher Scoring iterations: 5
```

```
> ivfinalmodel_noage = glm(case ~ tobacco + alcohol , family = binom
> ivfinalmodel_noage

Call:  glm(formula = case ~ tobacco + alcohol, family = binomial)

Coefficients:
(Intercept)      tobacco      alcohol
   -3.28618      0.02859      0.02497

Degrees of Freedom: 974 Total (i.e. Null);  972 Residual
Null Deviance:      989.5
Residual Deviance: 812.9        AIC: 818.9
> summary(ivfinalmodel_noage)

Call:
glm(formula = case ~ tobacco + alcohol, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8981  -0.6219  -0.4390  -0.2952   2.5511

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.286177   0.204890 -16.039  < 2e-16 ***
tobacco      0.028595   0.006758   4.231 2.32e-05 ***
alcohol      0.024968   0.002357  10.594  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 812.92  on 972  degrees of freedom
AIC: 818.92
```

```
> a = logLik(ivfinalmodel)
> b = logLik(ivfinalmodel_noage)
>
> G = -2*(b-(a))
> G
'log Lik.' 94.64345 (df=3)
> 1 - pchisq(G, 1)
'log Lik.' 0 (df=3)
```

Conclusion

- It is evident that alcohol, age, utilization of tobacco are in fact factors that influence the onset of Esophageal cancer.

```
> summary(ivfinalmodel)

Call:
glm(formula = case ~ age + tobacco + alcohol, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9323  -0.5923  -0.3206  -0.1410   2.8619

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.572920   0.586349 -12.915  < 2e-16 ***
age          0.072289   0.008222   8.792  < 2e-16 ***
tobacco      0.038803   0.007574   5.123 3.01e-07 ***
alcohol      0.026490   0.002525  10.492  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.49  on 974  degrees of freedom
Residual deviance: 718.28  on 971  degrees of freedom
AIC: 726.28

Number of Fisher Scoring iterations: 5
```

KEEP
CALM

AND

NO
QUESTIONS PLEASE