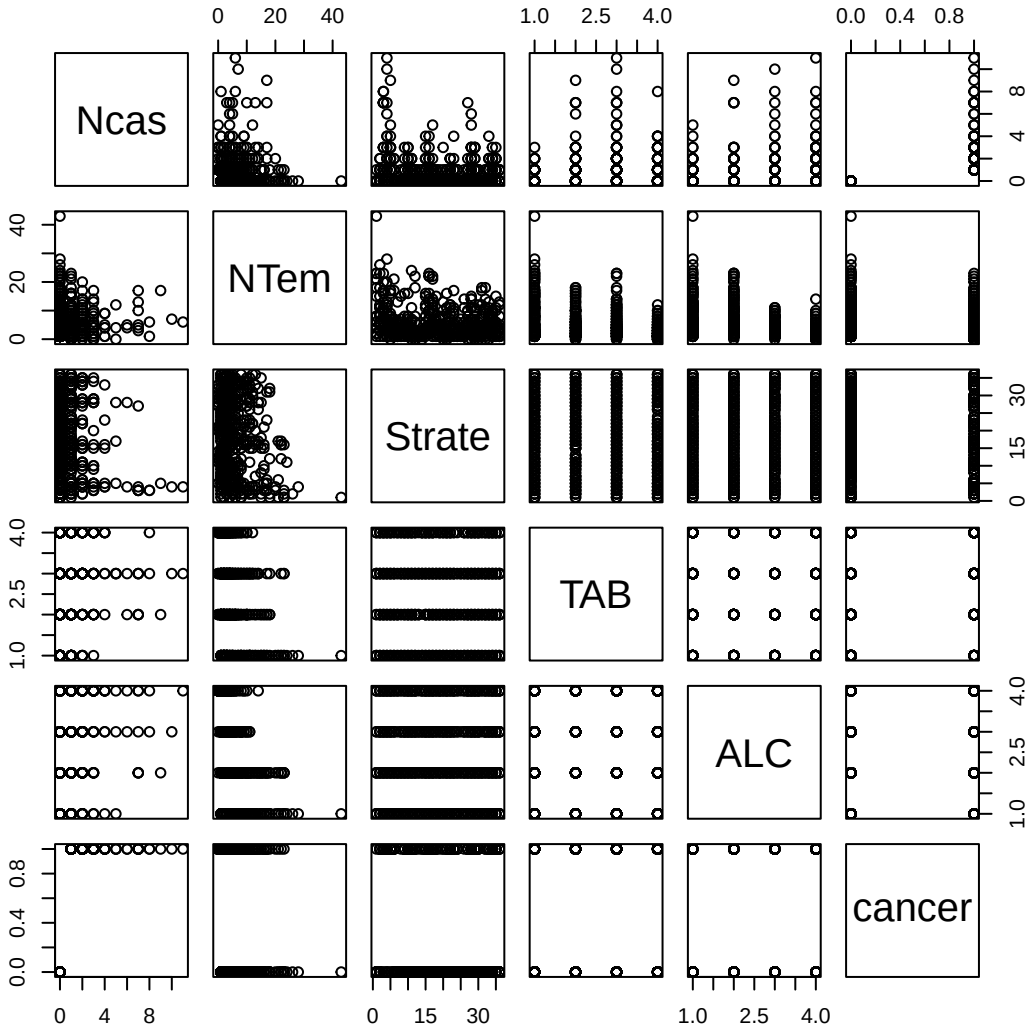
BE 2 - Régression Logistique

Question 1

```
Tuyns <- read.table("Tuyns.txt", header = TRUE)
Tuyns$TAB <- factor(Tuyns$TAB)
Tuyns$ALC <- factor(Tuyns$ALC)
summary(Tuyns)
```

```
##       Ncas              NTem            Strate      TAB     ALC
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 1.00   1:135   1:137
##  1st Qu.: 0.0000   1st Qu.: 2.000   1st Qu.: 9.00   2:128   2:136
##  Median : 0.0000   Median : 4.000   Median :18.00   3:137   3:122
##  Mean   : 0.7942   Mean   : 5.757   Mean   :18.43   4:115   4:120
##  3rd Qu.: 1.0000   3rd Qu.: 8.000   3rd Qu.:28.00
##  Max.   :11.0000   Max.   :43.000   Max.   :36.00
##      cancer
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3883
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
plot(Tuyns)
```




```r
reglTuyns <- glm(cancer ~ 1, family = binomial(link = "logit"), data = Tuyns)
summary(reglTuyns)
```

```
##
## Call:
## glm(formula = cancer ~ 1, family = binomial(link = "logit"),
##     data = Tuyns)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.45426    0.09041  -5.024 5.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 688.04  on 514  degrees of freedom
## Residual deviance: 688.04  on 514  degrees of freedom
## AIC: 690.04
##
## Number of Fisher Scoring iterations: 4
```

En régression logistique, p0 correspond à la probabilité de l'évènement Cancer = 1. Ainsi, alpha correspond à l'estimation de l'intercept moyen.

Si alpha est positif, la probabilité p0 est supérieure à 0,5, c'est-à-dire qu'il est plus probable qu'il y ait un cancer. Au contraire, si alpha est négatif, la probabilité est inférieure à 0,5.

donc ici alpha = -0,45 signifie que la probabilité d'avoir un cancer est plus faible que 0,5. La p-valeur associée à alpha est 5.06e-07 (très inférieure à 0.05). Cela signifie que la probabilité de base de survenue du cancer est significativement différente de 50%.

Question 2 :

On considère le modèle de régression logistique suivant :

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{TAB}_1 + \beta_2 \text{TAB}_2 + \beta_3 \text{TAB}_3$$

```r
model_tab <- glm(cancer ~ TAB, data = Tuyns, family = binomial)

anova(model_tab, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cancer
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  514     688.04
## TAB   3   36.549       511     651.50 5.732e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model_tab, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cancer
##     LR Chisq Df Pr(>Chisq)
## TAB   36.549  3  5.732e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model_tab)
```

```
##
## Call:
## glm(formula = cancer ~ TAB, family = binomial, data = Tuyns)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.4333     0.2183  -6.567 5.14e-11 ***
## TAB2          1.1182     0.2823   3.962 7.45e-05 ***
## TAB3          1.5648     0.2774   5.641 1.69e-08 ***
## TAB4          1.0639     0.2892   3.679 0.000234 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 688.04  on 514  degrees of freedom
## Residual deviance: 651.50  on 511  degrees of freedom
## AIC: 659.5
##
## Number of Fisher Scoring iterations: 4
```

L'intercept dans ce cas correspond au cas où TAB = 0.

Le Tabagisme est une variable influente, car sa p-value « 0,05.

Le résultat du summary nous indique que tous les niveaux d'exposition du tabagisme n'ont pas la même influence. Cependant ils ont tous une influence significative car leur p-value « 0,05.

Question 3 :

On considère le modèle de régression logistique suivant :

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{ALC}_1 + \beta_2 \text{ALC}_2 + \beta_3 \text{ALC}_3$$

```r
model_alc <- glm(cancer ~ ALC, data = Tuyns, family = binomial)

anova(model_alc, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cancer
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                  514     688.04
## ALC   3   37.202       511     650.84 4.17e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_alc, test.statistic = "LR", type = "III")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cancer
##     LR Chisq Df Pr(>Chisq)
## ALC   37.202  3   4.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model_alc)
```

```
##
## Call:
## glm(formula = cancer ~ ALC, family = binomial, data = Tuyns)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3591     0.2119  -6.415 1.41e-10 ***
## ALC2          0.9414     0.2750   3.424 0.000617 ***
## ALC3          0.9945     0.2807   3.543 0.000395 ***
## ALC4          1.6274     0.2808   5.796 6.77e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 688.04  on 514  degrees of freedom
## Residual deviance: 650.84  on 511  degrees of freedom
## AIC: 658.84
##
## Number of Fisher Scoring iterations: 4
```

```
anova(model_alc, model_tab)
```

```
## Analysis of Deviance Table
##
## Model 1: cancer ~ ALC
## Model 2: cancer ~ TAB
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       511     650.84
## 2       511     651.50  0 -0.65338
```

```
AIC(model_alc, model_tab)
```

```
##           df      AIC
## model_alc  4 658.8419
## model_tab  4 659.4953
```

L'intercept dans ce cas correspond au cas où ALC = 0.

L'e Tabagisme'Alcoolisme est une variable influente, car sa p-value « 0,05.

Le résultat du summary nous indique que tous les niveaux d'exposition d'u tabagisme d'alcoolisme n'ont pas la même influence. Cependant ils ont tous une influence significative car leur p-value « 0,05.

Les deux modèles ont des AIC très proches il est difficile de retenir une variable parmi les 2

Question 4 :

```r
# 1) Modèle complet avec interaction
model_interaction <- glm(cancer ~ TAB * ALC, data = Tuyns, family = binomial)
summary(model_interaction)
```

```
##
## Call:
## glm(formula = cancer ~ TAB * ALC, family = binomial, data = Tuyns)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.36712    0.60360  -3.922 8.79e-05 ***
## TAB2         1.18847    0.72650   1.636  0.10186
## TAB3         1.79658    0.69623   2.580  0.00987 **
## TAB4         0.42121    0.80626   0.522  0.60137
## ALC2         1.26851    0.71588   1.772  0.07640 .
## ALC3         0.82668    0.75281   1.098  0.27215
## ALC4         1.35552    0.73130   1.854  0.06380 .
## TAB2:ALC2   -0.37754    0.89029  -0.424  0.67152
## TAB3:ALC2   -0.58674    0.86275  -0.680  0.49645
## TAB4:ALC2    0.18492    0.97194   0.190  0.84910
## TAB2:ALC3    0.14434    0.93253   0.155  0.87699
## TAB3:ALC3    0.04925    0.90066   0.055  0.95639
## TAB4:ALC3    0.96508    1.00359   0.962  0.33624
## TAB2:ALC4    0.22860    0.91496   0.250  0.80271
## TAB3:ALC4    0.00348    0.89479   0.004  0.99690
## TAB4:ALC4    1.50668    0.99775   1.510  0.13102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 688.04  on 514  degrees of freedom
## Residual deviance: 606.25  on 499  degrees of freedom
## AIC: 638.25
##
## Number of Fisher Scoring iterations: 4
```

On remarque ici que les 2 seuls variables influentes sont l'intercept et TAB3.