TP3 - Exercice 1

EXPUESTO Ewen, AUBERTIN Candice

20 octobre 2025

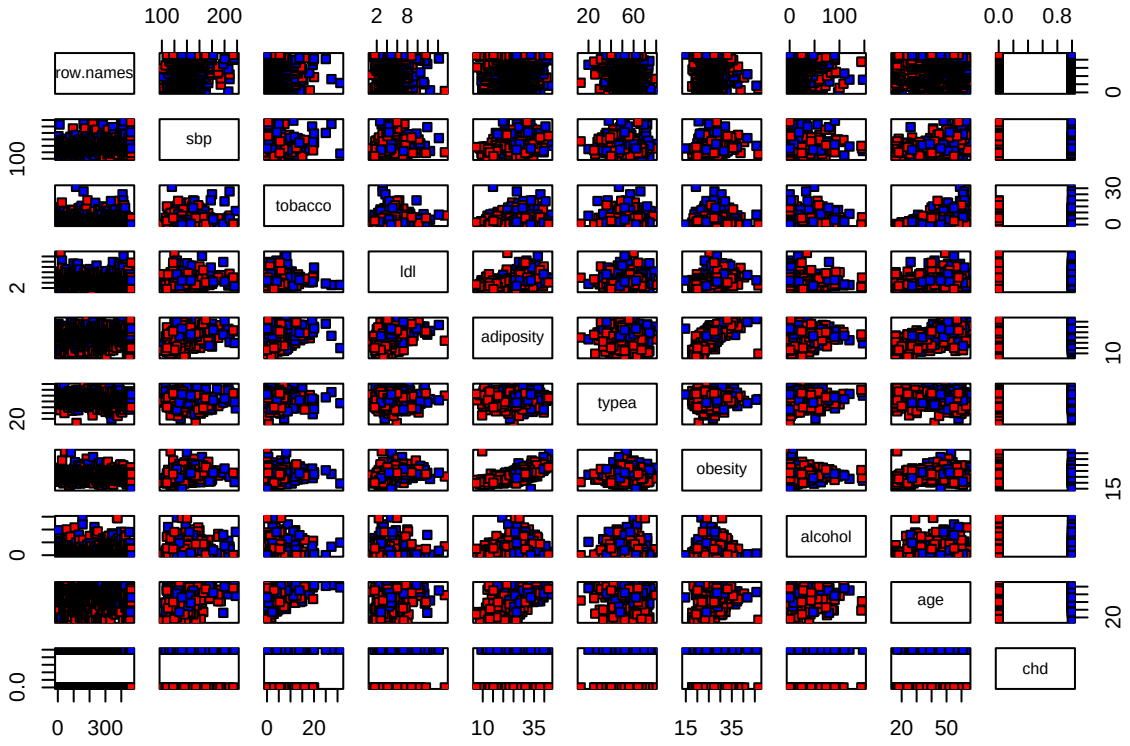
1 Introduction

```
rm(list = ls())
graphics.off()
```

```
saheart <- read.table("SAheart.txt", header = TRUE, sep = ",")
```

On enlève la colonne famhist, ou alors on pourrait la convertir en variable numérique.

```
saheart_num <- saheart[, sapply(saheart, is.numeric)]
pairs(saheart_num[], pch = 22, bg = c("red", "blue")[unclass(factor(saheart[, "chd"]))])
```



2 Modèle de régression logistique

2.1 a)

La fonction `glm()` est utilisée pour estimer les paramètres d'un modèle linéaire généralisé.

```
?glm
```

```
## starting httpd help server ... done
```

Explication de l'option `family = binomial`:

L'option `family = binomial` indique que nous souhaitons ajuster un modèle de régression logistique. Cette option spécifie :

- La distribution de la variable réponse (distribution binomiale pour une variable binaire 0/1)
- La fonction de lien (logit par défaut pour binomial), qui relie la moyenne de la variable réponse aux prédicteurs linéaires

Pour une régression logistique, le modèle est : $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$

où p est la probabilité que la variable réponse soit égale à 1 (présence de maladie coronarienne).

```
res <- glm(chd ~ ., data = saheart, family = binomial)
```

```
summary(res)
```

```
##
## Call:
## glm(formula = chd ~ ., family = binomial, data = saheart)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.9207616  1.3265724  -4.463 8.07e-06 ***
## row.names      -0.0008844  0.0008950  -0.988 0.323042
## sbp             0.0076602  0.0058574   1.308 0.190942
## tobacco         0.0777962  0.0266602   2.918 0.003522 **
## ldl             0.1701708  0.0597998   2.846 0.004432 **
## adiposity       0.0209609  0.0294496   0.712 0.476617
## famhistPresent  0.9385467  0.2287202   4.103 4.07e-05 ***
## typea           0.0376529  0.0124706   3.019 0.002533 **
## obesity        -0.0661926  0.0443180  -1.494 0.135285
## alcohol         0.0004222  0.0045053   0.094 0.925346
## age             0.0441808  0.0121784   3.628 0.000286 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 471.16  on 451  degrees of freedom
## AIC: 493.16
##
## Number of Fisher Scoring iterations: 5
```

2.2 b)

Affiche le contenu de l'objet res :

```
attributes(res)
```

```
## $names
##  [1] "coefficients"      "residuals"         "fitted.values"
##  [4] "effects"           "R"                 "rank"
##  [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
## [22] "call"              "formula"           "terms"
## [25] "data"              "offset"            "control"
## [28] "method"            "contrasts"         "xlevels"
##
## $class
## [1] "glm" "lm"
```

Affiche leur nom :

```
names(res)
```

```
##  [1] "coefficients"      "residuals"         "fitted.values"
##  [4] "effects"           "R"                 "rank"
##  [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
## [22] "call"              "formula"           "terms"
## [25] "data"              "offset"            "control"
## [28] "method"            "contrasts"         "xlevels"
```

Analyse des coefficients estimés:

Coefficients estimés :

```
res$coefficients
```

```
##    (Intercept)      row.names            sbp        tobacco            ldl
##  -5.9207615843  -0.0008844173   0.0076602312   0.0777961938   0.1701707546
##      adiposity famhistPresent          typea        obesity        alcohol
##   0.0209608624   0.9385466734   0.0376529123  -0.0661925542   0.0004221547
##            age
##   0.0441807735
```

```
coef_summary <- summary(res)$coefficients
coef_summary
```

```
##                     Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept)    -5.9207615843 1.326572408 -4.46320272 8.074359e-06
## row.names      -0.0008844173 0.000894955 -0.98822548 3.230422e-01
## sbp             0.0076602312 0.005857356  1.30779671 1.909423e-01
## tobacco         0.0777961938 0.026660229  2.91806177 3.522146e-03
## ldl             0.1701707546 0.059799842  2.84567233 4.431777e-03
## adiposity       0.0209608624 0.029449585  0.71175408 4.766171e-01
## famhistPresent  0.9385466734 0.228720162  4.10347152 4.069966e-05
## typea           0.0376529123 0.012470555  3.01934536 2.533216e-03
## obesity        -0.0661925542 0.044318009 -1.49358142 1.352851e-01
## alcohol         0.0004221547 0.004505290  0.09370201 9.253459e-01
## age             0.0441807735 0.012178434  3.62778763 2.858602e-04
```

Coefficients les plus significatifs:

Les coefficients les plus significatifs sont ceux avec les p-values les plus faibles (< 0.05). On peut les identifier en regardant la dernière colonne du tableau ci-dessus.

```
significant_coef <- coef_summary[coef_summary[, 4] < 0.05, ]
significant_coef
```

```
##                   Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)    -5.92076158 1.32657241 -4.463203 8.074359e-06
## tobacco         0.07779619 0.02666023  2.918062 3.522146e-03
## ldl             0.17017075 0.05979984  2.845672 4.431777e-03
## famhistPresent  0.93854667 0.22872016  4.103472 4.069966e-05
## typea           0.03765291 0.01247055  3.019345 2.533216e-03
## age             0.04418077 0.01217843  3.627788 2.858602e-04
```

Coefficients les moins significatifs:

```
non_significant_coef <- coef_summary[coef_summary[, 4] >= 0.05, ]
non_significant_coef
```

```
##                Estimate  Std. Error     z value  Pr(>|z|)
## row.names -0.0008844173 0.000894955 -0.98822548 0.3230422
## sbp        0.0076602312 0.005857356  1.30779671 0.1909423
## adiposity  0.0209608624 0.029449585  0.71175408 0.4766171
## obesity   -0.0661925542 0.044318009 -1.49358142 0.1352851
## alcohol    0.0004221547 0.004505290  0.09370201 0.9253459
```

Prédictions avec predict():

Prédictions avec type = “link” (échelle logit) :

```
pred_link <- predict(res, type = "link")
head(pred_link)
```

```
##          1          2          3          4          5          6
##  1.1443307 -0.5243464 -0.7429019  1.1634194  1.0036471  0.6578029
```

Prédictions avec type = “response” (probabilités) :

```r
pred_response <- predict(res, type = "response")
head(pred_response)
```

```
##         1         2         3         4         5         6
## 0.7584739 0.3718365 0.3223699 0.7619535 0.7317750 0.6587667
```

Différence entre type = "link" et type = "response":

- `type = "link"` : retourne les valeurs prédites sur l'échelle du lien logit, c'est-à-dire $\hat{\beta}^T x_i = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$
- `type = "response"` : retourne les probabilités prédites, c'est-à-dire $\hat{p}_i = \frac{1}{1+e^{-\hat{\beta}^T x_i}}$

Vérification de la relation entre les deux :

```r
head(exp(pred_link) / (1 + exp(pred_link)))
```

```
##         1         2         3         4         5         6
## 0.7584739 0.3718365 0.3223699 0.7619535 0.7317750 0.6587667
```

```r
head(pred_response)
```

```
##         1         2         3         4         5         6
## 0.7584739 0.3718365 0.3223699 0.7619535 0.7317750 0.6587667
```

2.3 c) Odds-ratios

L'**odds-ratio** pour une variable X_j est calculé comme $OR_j = e^{\beta_j}$. Il représente le facteur multiplicatif du rapport des cotes (odds) lorsque X_j augmente d'une unité, toutes les autres variables restant constantes.

Calcul des odds-ratios :

```r
odds_ratios <- exp(res$coefficients)
odds_ratios
```

```
##    (Intercept)      row.names            sbp        tobacco            ldl
##    0.002683156    0.999115974    1.007689646    1.080902342    1.185507265
##      adiposity famhistPresent          typea        obesity        alcohol
##    1.021182084    2.556263631    1.038370765    0.935950626    1.000422244
##            age
##    1.045171277
```

Intervalles de confiance à 95% pour les odds-ratios :

```r
odds_ratios_ci <- exp(confint(res))
```

```
## Waiting for profiling to be done...
```

```
odds_ratios_ci
```

```
##                       2.5 %     97.5 %
## (Intercept)    0.0001875832 0.03443924
## row.names      0.9973517540 1.00086423
## sbp            0.9962425025 1.01946605
## tobacco        1.0271671548 1.14075639
## ldl            1.0566392105 1.33689964
## adiposity      0.9642716622 1.08261657
## famhistPresent 1.6368995221 4.01791338
## typea          1.0137282787 1.06462471
## obesity        0.8559190974 1.01902332
## alcohol        0.9915117000 1.00930476
## age            1.0207848474 1.07082656
```

Tableau complet avec odds-ratios et IC :

```
odds_table <- data.frame(
  Coefficient = res$coefficients,
  OddsRatio = odds_ratios,
  CI_lower = odds_ratios_ci[, 1],
  CI_upper = odds_ratios_ci[, 2]
)
odds_table
```

```
##                  Coefficient   OddsRatio     CI_lower   CI_upper
## (Intercept)    -5.9207615843 0.002683156 0.0001875832 0.03443924
## row.names      -0.0008844173 0.999115974 0.9973517540 1.00086423
## sbp             0.0076602312 1.007689646 0.9962425025 1.01946605
## tobacco         0.0777961938 1.080902342 1.0271671548 1.14075639
## ldl             0.1701707546 1.185507265 1.0566392105 1.33689964
## adiposity       0.0209608624 1.021182084 0.9642716622 1.08261657
## famhistPresent  0.9385466734 2.556263631 1.6368995221 4.01791338
## typea           0.0376529123 1.038370765 1.0137282787 1.06462471
## obesity        -0.0661925542 0.935950626 0.8559190974 1.01902332
## alcohol         0.0004221547 1.000422244 0.9915117000 1.00930476
## age             0.0441807735 1.045171277 1.0207848474 1.07082656
```

Interprétation de l'odds-ratio pour "tobacco":

Odds-ratio pour tobacco :

```
OR_tobacco <- odds_ratios["tobacco"]
OR_tobacco
```

```
##  tobacco
## 1.080902
```

Commentaire: Un odds-ratio de 1.081 pour la variable "tobacco" signifie que pour chaque kilogramme supplémentaire de tabac consommé (cumulative), le rapport des cotes de développer une maladie coronarienne est multiplié par 1.081, toutes les autres variables restant constantes.

Limites de cette approche:

1. Les odds-ratios peuvent être difficiles à interpréter intuitivement
2. L'interprétation suppose que toutes les autres variables restent constantes (ceteris paribus)
3. Pour les variables continues, l'effet d'une augmentation d'une unité peut ne pas être cliniquement pertinent
4. Les odds-ratios ne sont pas des risques relatifs (bien qu'ils s'en rapprochent pour les événements rares)
5. L'interprétation assume une relation linéaire entre le prédicteur et le log-odds