Survival data analysis - Lab2

Jeremy Bron

10 décembre 2018

EXO 1

Construction and interpretation of a Cox model for the pharmocoSmoking dataset I

Question 1 : How many covariates does the dataset contain ?

```
data("pharmacoSmoking")
glimpse(pharmacoSmoking)
```

```
## Observations: 125
## Variables: 14
## $ id             <int> 21, 113, 39, 80, 87, 29, 16, 35, 54, 70, 84, 85...
## $ ttr            <int> 182, 14, 5, 16, 0, 182, 14, 77, 2, 0, 12, 182, ...
## $ relapse        <int> 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,...
## $ grp            <fct> patchOnly, patchOnly, combination, combination,...
## $ age            <int> 36, 41, 25, 54, 45, 43, 66, 78, 40, 38, 64, 51,...
## $ gender         <fct> Male, Male, Female, Male, Male, Male, Male, Fem...
## $ race           <fct> white, white, white, white, white, hispanic, bl...
## $ employment     <fct> ft, other, other, ft, other, ft, pt, other, ft,...
## $ yearsSmoking   <int> 26, 27, 12, 39, 30, 30, 54, 56, 25, 23, 30, 35,...
## $ levelSmoking   <fct> heavy, heavy, heavy, heavy, heavy, heavy, heavy...
## $ ageGroup2      <fct> 21-49, 21-49, 21-49, 50+, 21-49, 21-49, 50+, 50...
## $ ageGroup4      <fct> 35-49, 35-49, 21-34, 50-64, 35-49, 35-49, 65+, ...
## $ priorAttempts  <int> 0, 3, 3, 0, 0, 2, 0, 10, 4, 10, 12, 1, 5, 6, 5,...
## $ longestNoSmoke <int> 0, 90, 21, 0, 0, 1825, 0, 15, 7, 90, 365, 7, 10...
```

```
#Les transformations ci-après sont réalisés pour corriger des individues abérants vu en cour
s.
pharmacoSmoking[pharmacoSmoking$longestNoSmoke>4000,]$longestNoSmoke=4000
pharmacoSmoking[pharmacoSmoking$priorAttempts>200,]$priorAttempts=150
```

Le jeu de données contient 14 colonnes. Les variables ageGroup2 et ageGroup4 indiquent dans quelle tranche d'âge se trouve l'individu selon un découpage en 2 ou 4 intervalles. Ainsi les informations de ageGroup2 sont redondantes, car contenues dans ageGroup4. Ainsi on a en réalité 13 covariables. La variable ID doit aussi être retiré, car elle n'apporte pas d'informations.

À noter que les variables catégorielles à plus de 2 facteurs sont séparées en plusieurs covariables lors des fits ce qui donne 16 covariables lorsque l'on exécute le modèle complet (avec toutes les variables sauf ID).

Question 2 : Fit a first Cox model with all the covariates you found in question 1. What is the problem ?

Pour cette question on utilise toutes les variables du dataset.

```
summary(coxph(Surv(ttr,relapse) ~ . - id, data = pharmacoSmoking))
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ . - id, data = pharmacoSmoking)
##
##   n= 125, number of events= 89
##
##                        coef  exp(coef)   se(coef)      z Pr(>|z|)
## grppatchOnly       0.6062913  1.8336184  0.2213496  2.739  0.00616 **
## age               -0.0412217  0.9596164  0.0316335 -1.303  0.19254
## genderMale        -0.0209507  0.9792672  0.2563696 -0.082  0.93487
## racehispanic      -0.4695080  0.6253099  0.5292019 -0.887  0.37497
## raceother         -1.1801998  0.3072174  1.0534418 -1.120  0.26257
## racewhite         -0.2989356  0.7416072  0.2648844 -1.129  0.25909
## employmentother    0.7277932  2.0705064  0.2896467  2.513  0.01198 *
## employmentpt       0.5995188  1.8212422  0.3443748  1.741  0.08170 .
## yearsSmoking       0.0129266  1.0130105  0.0193901  0.667  0.50499
## levelSmokinglight -0.0482579  0.9528880  0.2882013 -0.167  0.86702
## ageGroup250+       0.3701667  1.4479759  1.2185055  0.304  0.76129
## ageGroup435-49     0.2029720  1.2250382  0.4842957  0.419  0.67514
## ageGroup450-64    -0.6796180  0.5068106  0.6014610 -1.130  0.25850
## ageGroup465+              NA         NA  0.0000000     NA       NA
## priorAttempts     -0.0050533  0.9949595  0.0084302 -0.599  0.54889
## longestNoSmoke    -0.0001220  0.9998780  0.0001375 -0.887  0.37489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly         1.8336     0.5454   1.18822     2.830
## age                  0.9596     1.0421   0.90193     1.021
## genderMale           0.9793     1.0212   0.59249     1.619
## racehispanic         0.6253     1.5992   0.22163     1.764
## raceother            0.3072     3.2550   0.03897     2.422
## racewhite            0.7416     1.3484   0.44127     1.246
## employmentother      2.0705     0.4830   1.17362     3.653
## employmentpt         1.8212     0.5491   0.92733     3.577
## yearsSmoking         1.0130     0.9872   0.97523     1.052
## levelSmokinglight    0.9529     1.0494   0.54166     1.676
## ageGroup250+         1.4480     0.6906   0.13291    15.774
## ageGroup435-49       1.2250     0.8163   0.47415     3.165
## ageGroup450-64       0.5068     1.9731   0.15591     1.647
## ageGroup465+             NA         NA        NA        NA
## priorAttempts        0.9950     1.0051   0.97865     1.012
## longestNoSmoke       0.9999     1.0001   0.99961     1.000
##
## Concordance= 0.676  (se = 0.034 )
## Rsquare= 0.23   (max possible= 0.998 )
## Likelihood ratio test= 32.65  on 15 df,   p=0.005
## Wald test            = 30.41  on 15 df,   p=0.01
## Score (logrank) test = 31.59  on 15 df,   p=0.007
```

On constate que le modèle n'a pas pu calcules les coefficients et tests pour ageGroup465+. Le problème réside dans la redondance entre ageGroup4 et ageGroup2. Il existe ici une relation linéaire parfaite entre ces deux variables.

Ainsi le calcul des coefficients ne peut se faire pour la covariable ageGroup465+ et renvoi NA.

Question 3 : Fit a new Cox model with a subset of covariates, that solves the previous problem.

Afin de résoudre le problème, on utilise un sous-ensemble de covariables qui ne contient pas ageGroup2.

```
mod1 = coxph(Surv(ttr,relapse) ~ . - ageGroup2 - id, data = pharmacoSmoking)
summary(mod1)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ . - ageGroup2 - id, data = pharmacoSmoking)
##
##   n= 125, number of events= 89
##
##                        coef  exp(coef)   se(coef)      z Pr(>|z|)
## grppatchOnly       0.6062913  1.8336184  0.2213496  2.739  0.00616 **
## age               -0.0412217  0.9596164  0.0316335 -1.303  0.19254
## genderMale        -0.0209507  0.9792672  0.2563696 -0.082  0.93487
## racehispanic      -0.4695080  0.6253099  0.5292019 -0.887  0.37497
## raceother         -1.1801998  0.3072174  1.0534418 -1.120  0.26257
## racewhite         -0.2989356  0.7416072  0.2648844 -1.129  0.25909
## employmentother    0.7277932  2.0705064  0.2896467  2.513  0.01198 *
## employmentpt       0.5995188  1.8212422  0.3443748  1.741  0.08170 .
## yearsSmoking       0.0129266  1.0130105  0.0193901  0.667  0.50499
## levelSmokinglight -0.0482579  0.9528880  0.2882013 -0.167  0.86702
## ageGroup435-49     0.2029720  1.2250382  0.4842957  0.419  0.67514
## ageGroup450-64    -0.3094513  0.7338495  0.7887222 -0.392  0.69480
## ageGroup465+       0.3701667  1.4479759  1.2185055  0.304  0.76129
## priorAttempts     -0.0050533  0.9949595  0.0084302 -0.599  0.54889
## longestNoSmoke    -0.0001220  0.9998780  0.0001375 -0.887  0.37489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly         1.8336     0.5454   1.18822     2.830
## age                  0.9596     1.0421   0.90193     1.021
## genderMale           0.9793     1.0212   0.59249     1.619
## racehispanic         0.6253     1.5992   0.22163     1.764
## raceother            0.3072     3.2550   0.03897     2.422
## racewhite            0.7416     1.3484   0.44127     1.246
## employmentother      2.0705     0.4830   1.17362     3.653
## employmentpt         1.8212     0.5491   0.92733     3.577
## yearsSmoking         1.0130     0.9872   0.97523     1.052
## levelSmokinglight    0.9529     1.0494   0.54166     1.676
## ageGroup435-49       1.2250     0.8163   0.47415     3.165
## ageGroup450-64       0.7338     1.3627   0.15640     3.443
## ageGroup465+         1.4480     0.6906   0.13291    15.774
## priorAttempts        0.9950     1.0051   0.97865     1.012
## longestNoSmoke       0.9999     1.0001   0.99961     1.000
##
## Concordance= 0.676  (se = 0.034 )
## Rsquare= 0.23   (max possible= 0.998 )
## Likelihood ratio test= 32.65  on 15 df,   p=0.005
## Wald test            = 30.41  on 15 df,   p=0.01
## Score (logrank) test = 31.59  on 15 df,   p=0.007
```

Question 4 : Do a backward procedure of variable selection based on Wald tests.

```
# pour avoir le BIC, mettre k = log (n)
# pour avoir le AIC, mettre k = 2

#Step dans la direction backward
step(mod1, direction = "backward", k=log(125),trace=0 )
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + age, data = pharmacoSmoking)
##
##                 coef exp(coef) se(coef)     z      p
## grppatchOnly  0.5587    1.7483   0.2167  2.58 0.0099
## age          -0.0230    0.9772   0.0096 -2.40 0.0165
##
## Likelihood ratio test=13.82  on 2 df, p=0.001
## n= 125, number of events= 89
```

```
step(mod1, direction = "backward", k=2,trace=0)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + age + employment,
##     data = pharmacoSmoking)
##
##                    coef exp(coef) se(coef)     z      p
## grppatchOnly     0.6079    1.8365   0.2184  2.78 0.0054
## age              -0.0353    0.9653   0.0108 -3.28 0.0010
## employmentother  0.7035    2.0208   0.2693  2.61 0.0090
## employmentpt     0.6537    1.9226   0.3273  2.00 0.0458
##
## Likelihood ratio test=22.03  on 4 df, p=2e-04
## n= 125, number of events= 89
```

La direction du step (backward ou both) ne change pas la sélection des variables. Avec le BIC on pénalise plus et il reste seulement 2 variables. Pour la suite on gardera seulement le critère AIC. La fonction step à éliminer le groupe d'âge ageGroup2, dès lors, la variable âge qui est linéaire va avoir une influence similaire sur toute la plage d'âge, ce qui est faux: une augmentation d'un an à 20 ans n'est pas pareil que 1 an à 60 ans…!

Pour résoudre ce problème, on peut:

- Enlever la variable âge: step sur le modèle sans age
- rajouter manuellement agegroupe4 au modèle précédent

```
mod2 = coxph(Surv(ttr,relapse) ~ . - ageGroup2 - id - age , data = pharmacoSmoking)
mod2s = step(mod2, direction = "backward", k=2,trace=0)
summary(mod2s)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + employment + ageGroup4,
##     data = pharmacoSmoking)
##
##   n= 125, number of events= 89
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## grppatchOnly     0.6564    1.9278   0.2198  2.986  0.00283 **
## employmentother  0.6231    1.8648   0.2764  2.254  0.02418 *
## employmentpt     0.5214    1.6844   0.3320  1.570  0.11631
## ageGroup435-49  -0.1119    0.8942   0.3216 -0.348  0.72792
## ageGroup450-64  -1.0233    0.3594   0.3597 -2.845  0.00444 **
## ageGroup465+    -0.7071    0.4931   0.5017 -1.410  0.15868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly       1.9278     0.5187    1.2529    2.9661
## employmentother    1.8648     0.5363    1.0848    3.2057
## employmentpt       1.6844     0.5937    0.8787    3.2289
## ageGroup435-49     0.8942     1.1184    0.4761    1.6793
## ageGroup450-64     0.3594     2.7825    0.1776    0.7273
## ageGroup465+       0.4931     2.0281    0.1845    1.3180
##
## Concordance= 0.647  (se = 0.034 )
## Rsquare= 0.187   (max possible= 0.998 )
## Likelihood ratio test= 25.89  on 6 df,   p=2e-04
## Wald test            = 24.59  on 6 df,   p=4e-04
## Score (logrank) test = 25.54  on 6 df,   p=3e-04
```

```
mod3 = coxph(Surv(ttr,relapse) ~ grp + age + ageGroup4 + employment , data = pharmacoSmoking)
summary(mod3)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + age + ageGroup4 +
##     employment, data = pharmacoSmoking)
##
##   n= 125, number of events= 89
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## grppatchOnly     0.62992   1.87745  0.22086  2.852  0.00434 **
## age             -0.03563   0.96500  0.02616 -1.362  0.17319
## ageGroup435-49   0.29562   1.34395  0.43856  0.674  0.50027
## ageGroup450-64  -0.17831   0.83668  0.71655 -0.249  0.80348
## ageGroup465+     0.65842   1.93174  1.11870  0.589  0.55616
## employmentother  0.68535   1.98447  0.27867  2.459  0.01392 *
## employmentpt     0.58170   1.78907  0.33282  1.748  0.08050 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly       1.8775     0.5326    1.2178     2.894
## age                0.9650     1.0363    0.9168     1.016
## ageGroup435-49     1.3440     0.7441    0.5690     3.175
## ageGroup450-64     0.8367     1.1952    0.2054     3.408
## ageGroup465+       1.9317     0.5177    0.2156    17.305
## employmentother    1.9845     0.5039    1.1493     3.426
## employmentpt       1.7891     0.5589    0.9318     3.435
##
## Concordance= 0.653  (se = 0.034 )
## Rsquare= 0.199   (max possible= 0.998 )
## Likelihood ratio test= 27.76  on 7 df,   p=2e-04
## Wald test            = 26.55  on 7 df,   p=4e-04
## Score (logrank) test = 27.5  on 7 df,   p=3e-04
```

On conserve le modèle mod2s.

Question 4 : Interpret the coefficients in the final model.

```
mod2 = coxph(Surv(ttr,relapse) ~ . - ageGroup2 - id - age , data = pharmacoSmoking)
mod2s = step(mod2, direction = "backward", k=2,trace=0)
summary(mod2s)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + employment + ageGroup4,
##     data = pharmacoSmoking)
##
##   n= 125, number of events= 89
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## grppatchOnly      0.6564    1.9278   0.2198  2.986  0.00283 **
## employmentother   0.6231    1.8648   0.2764  2.254  0.02418 *
## employmentpt      0.5214    1.6844   0.3320  1.570  0.11631
## ageGroup435-49   -0.1119    0.8942   0.3216 -0.348  0.72792
## ageGroup450-64   -1.0233    0.3594   0.3597 -2.845  0.00444 **
## ageGroup465+     -0.7071    0.4931   0.5017 -1.410  0.15868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly       1.9278     0.5187    1.2529    2.9661
## employmentother    1.8648     0.5363    1.0848    3.2057
## employmentpt       1.6844     0.5937    0.8787    3.2289
## ageGroup435-49     0.8942     1.1184    0.4761    1.6793
## ageGroup450-64     0.3594     2.7825    0.1776    0.7273
## ageGroup465+       0.4931     2.0281    0.1845    1.3180
##
## Concordance= 0.647  (se = 0.034 )
## Rsquare= 0.187   (max possible= 0.998 )
## Likelihood ratio test= 25.89  on 6 df,   p=2e-04
## Wald test            = 24.59  on 6 df,   p=4e-04
## Score (logrank) test = 25.54  on 6 df,   p=3e-04
```

Interprétation des coefficients:

- grppatchOnly: utilisation du patch seul (au lieu de patch + autre chose): exp(coef) = 1.9278 : le risque de rechute est multiplié par 1,93. C'est à dire deux fois plus de rechute si l'on utilise seulement le patch. Cette variable est la plus significative (p value = 0.00283), on peut donc inférerez sur l'effet du traitement sur la rechute avec beaucoup de certitudes
- employment: cette variable qualitative a 3 niveaux: ft (full time job), pt (part time job), other. Les individus pt ou other ont respectivement un risque de rechute multiplié par 1.6844 et 1.8648 comparé à un individue ft. Le coefficient pour other est significatif (p value = 0.02418); celui de pt est relativement significatif également (p value = 0.11631).Globalement les individus qui ne travaillent pas à temps plein ont un risque de rechute multiplié par environ 1,75.
- ageGroup4: sépare l'âge des individus en 4 groupes (21-34, 35-49, 50-64, 65+). Les coefficients des groupes 35-49, 50-64, 65+ sont tous négatifs (l'exponention est inférieur à 1) : le risque de rechute diminue globalement pour les individus de moins de 34 ans. En détail:
    - ageGroup435-49: la p value est élevé (0.72792) et le exp(coef) proche de 0,9. l'appartenance à ce groupe n'est pas significative sur le risque de rechute.
    - ageGroup450-64: la p value est faible (0.00444) et le exp(coef) = 0.3594. L'appartenance à ce groupe d'âge influence significativement le risque de rechute à la baisse: il est multiplié par 0.3594.
    - ageGroup465+: la p value est un peu élevé (0.15868) et le exp(coef) = 0.4931. L'appartenance à ce groupe d'âge influence le risque de rechute à la baisse: il est globalement divisé par 2.