Time-consuming Calculation

Hanna Nguyen   Nathaniel Shek Wing Cheng

4/16/2020

```
rmarkdown::render("time_consuming_calculation.Rmd", params = list(load_calcs = FALSE))
```

1 Code running slow for choosing candidate model

```
# intercept only
M0 <- lm(log(chdrisk) - log(1 - chdrisk) ~ 1, data = fhs)
# all main effects and interactions
Mmax <- lm(log(chdrisk) - log(1 - chdrisk) ~ (.)^2, data = fhs)

names(beta.max)[is.na(beta.max)]
Mmax <- lm(log(chdrisk) - log(1 - chdrisk) ~ (.)^2 - cursmoke:cigpday -
             bpmeds:prevhyp, data = fhs)

# starting point model: main effects only
Mstart <- lm(log(chdrisk) - log(1 - chdrisk) ~ ., data = fhs)

# forward
system.time({
  Mfwd <- step(object = M0, # base model
               scope = list(lower = M0, upper = Mmax), # smallest and largest model
               direction = "forward",
               trace = FALSE) # trace prints out information
})

# backward
system.time({
  Mback <- step(object = Mmax, # base model
                scope = list(lower = M0, upper = Mmax),
                direction = "backward", trace = FALSE)
})

# stepwise
system.time({
  Mstep <- step(object = Mstart,
                scope = list(lower = M0, upper = Mmax),
                direction = "both", trace = FALSE)
})
```

```
# Second Candidate Model
M1 <- lm(log(chdrisk) - log(1 - chdrisk) ~ ., data = fhs)
M2 <- lm(log(chdrisk) - log(1 - chdrisk) ~ . - totchol - ldlc, data = fhs)
```

1.1 Summary of first candidate model

```
summary(Mstep)
```

```
##
## Call:
## lm(formula = log(chdrisk) - log(1 - chdrisk) ~ sex + totchol +
##     age + sysbp + diabp + cursmoke + cigpday + bmi + diabetes +
##     bpmeds + heartrte + glucose + prevmi + prevstrk + prevhyp +
##     hdlc + ldlc + sysbp:prevmi + age:diabp + totchol:prevhyp +
##     totchol:hdlc + hdlc:ldlc + diabetes:prevmi + prevhyp:ldlc +
##     sysbp:diabetes + totchol:heartrte + sysbp:prevhyp + sysbp:diabp +
##     bmi:ldlc + prevmi:prevhyp + sysbp:heartrte + sex:glucose +
##     age:cigpday + prevmi:hdlc + sysbp:hdlc + age:ldlc + sex:sysbp +
##     prevmi:ldlc + age:heartrte + sysbp:bpmeds + sysbp:cursmoke +
##     age:glucose + diabp:prevhyp + age:prevmi + bmi:prevmi + diabetes:hdlc +
##     cigpday:hdlc + cursmoke:hdlc + age:prevhyp + cursmoke:ldlc +
##     sex:totchol + prevmi:prevstrk + cigpday:glucose + diabp:bpmeds +
##     totchol:ldlc + bmi:bpmeds + sex:prevhyp + cursmoke:bpmeds,
##     data = fhs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67803 -0.28338  0.01316  0.28937  1.72607
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -7.716e+00  1.050e+00  -7.347 2.82e-13 ***
## sexMale                 8.495e-01  1.951e-01   4.354 1.40e-05 ***
## totchol                -1.794e-03  2.211e-03  -0.812 0.417114
## age                     7.599e-02  1.255e-02   6.057 1.62e-09 ***
## sysbp                   3.588e-03  5.028e-03   0.714 0.475592
## diabp                  -1.665e-02  1.012e-02  -1.645 0.100018
## cursmokeYes             1.789e-01  2.134e-01   0.838 0.402028
## cigpday                 4.030e-02  9.103e-03   4.427 1.00e-05 ***
## bmi                    -2.165e-02  1.109e-02  -1.953 0.050948 .
## diabetesYes             1.155e+00  2.853e-01   4.048 5.33e-05 ***
## bpmedsYes               7.494e-01  3.268e-01   2.293 0.021919 *
## heartrte                4.948e-02  7.831e-03   6.319 3.17e-10 ***
## glucose                -2.281e-03  2.861e-03  -0.797 0.425341
## prevmiYes               6.267e+00  6.166e-01  10.164  < 2e-16 ***
## prevstrkYes             1.829e-01  8.055e-02   2.271 0.023256 *
## prevhypYes              3.649e+00  4.043e-01   9.024  < 2e-16 ***
## hdlc                   -2.759e-02  5.804e-03  -4.753 2.13e-06 ***
## ldlc                   -5.422e-04  2.916e-03  -0.186 0.852479
## sysbp:prevmiYes        -8.676e-03  2.615e-03  -3.318 0.000920 ***
## age:diabp              -4.557e-04  1.295e-04  -3.518 0.000444 ***
## totchol:prevhypYes    -6.171e-03  1.227e-03  -5.031 5.27e-07 ***
## totchol:hdlc           3.009e-04  2.055e-05  14.642  < 2e-16 ***
## hdlc:ldlc             -2.441e-04  1.884e-05 -12.954  < 2e-16 ***
## diabetesYes:prevmiYes -6.662e-01  1.349e-01  -4.937 8.50e-07 ***
## prevhypYes:ldlc        3.005e-03  1.183e-03   2.541 0.011132 *  
## sysbp:diabetesYes     -6.780e-03  1.655e-03  -4.097 4.33e-05 ***
## totchol:heartrte      -6.721e-05  1.906e-05  -3.527 0.000429 ***
## sysbp:prevhypYes      -8.953e-03  2.056e-03  -4.354 1.40e-05 ***
## sysbp:diabp            3.543e-04  5.274e-05   6.719 2.32e-11 ***
## bmi:ldlc               2.043e-04  5.910e-05   3.458 0.000555 ***
## prevmiYes:prevhypYes  -3.686e-01  1.336e-01  -2.759 0.005852 ** 
## sysbp:heartrte        -1.140e-04  3.804e-05  -2.996 0.002765 ** 
## sexMale:glucose       -2.060e-03  7.073e-04  -2.912 0.003623 ** 
## age:cigpday           -2.890e-04  1.294e-04  -2.234 0.025609 *  
## prevmiYes:hdlc         1.117e-02  3.819e-03   2.926 0.003470 ** 
## sysbp:hdlc            -1.075e-04  3.442e-05  -3.123 0.001816 ** 
## age:ldlc               6.035e-05  2.883e-05   2.093 0.036460 *  
## sexMale:sysbp         -3.680e-03  1.246e-03  -2.955 0.003164 ** 
## prevmiYes:ldlc        -2.863e-03  9.315e-04  -3.074 0.002139 ** 
## age:heartrte          -2.506e-04  1.011e-04  -2.478 0.013279 *  
## sysbp:bpmedsYes       -5.752e-03  1.777e-03  -3.237 0.001227 ** 
## sysbp:cursmokeYes     -1.553e-03  1.099e-03  -1.413 0.157730    
## age:glucose            8.069e-05  4.427e-05   1.822 0.068523 .  
## diabp:prevhypYes      -1.030e-02  3.752e-03  -2.746 0.006077 ** 
## age:prevmiYes         -1.162e-02  6.546e-03  -1.775 0.076019 .  
## bmi:prevmiYes         -2.119e-02  1.243e-02  -1.705 0.088319 .  
## diabetesYes:hdlc       5.755e-03  2.447e-03   2.351 0.018790 *  
## cigpday:hdlc          -3.008e-04  9.572e-05  -3.142 0.001699 ** 
## cursmokeYes:hdlc       5.344e-03  2.370e-03   2.254 0.024268 *  
## age:prevhypYes        -6.835e-03  3.386e-03  -2.018 0.043663 *  
## cursmokeYes:ldlc      -7.778e-04  4.990e-04  -1.559 0.119175    
## sexMale:totchol        1.065e-03  5.156e-04   2.066 0.038934 *  
## prevmiYes:prevstrkYes -3.239e-01  1.986e-01  -1.631 0.102944    
## cigpday:glucose       -6.675e-05  3.793e-05  -1.760 0.078604 .  
## diabp:bpmedsYes        7.124e-03  3.320e-03   2.146 0.031988 *  
## totchol:ldlc           4.298e-06  2.516e-06   1.709 0.087671 .  
## bmi:bpmedsYes         -1.334e-02  7.177e-03  -1.858 0.063265 .  
## sexMale:prevhypYes     8.497e-02  5.633e-02   1.509 0.131553    
## cursmokeYes:bpmedsYes -1.049e-01  6.975e-02  -1.504 0.132625    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4764 on 2247 degrees of freedom
## Multiple R-squared:  0.8447, Adjusted R-squared:  0.8407 
## F-statistic: 210.7 on 58 and 2247 DF,  p-value: < 2.2e-16
```

1.2 Summary of second candidate model

```r
summary(M2)
```

```
## 
## Call:
```

```
## lm(formula = log(chdrisk) - log(1 - chdrisk) ~ . - totchol -
##     ldlc, data = fhs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90510 -0.32676  0.00583  0.37289  2.30109
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.6358726  0.1843492 -30.572  < 2e-16 ***
## sexMale      0.4259739  0.0261001  16.321  < 2e-16 ***
## age          0.0346995  0.0017384  19.961  < 2e-16 ***
## sysbp        0.0077335  0.0009059   8.537  < 2e-16 ***
## diabp        0.0019784  0.0016216   1.220  0.22259
## cursmokeYes  0.0161609  0.0429762   0.376  0.70692
## cigpday      0.0047942  0.0017663   2.714  0.00669 **
## bmi          0.0182133  0.0033069   5.508 4.04e-08 ***
## diabetesYes  0.3275994  0.0523025   6.264 4.48e-10 ***
## bpmedsYes    0.1069385  0.0372304   2.872  0.00411 **
## heartrte     0.0030825  0.0010029   3.073  0.00214 **
## glucose      0.0013770  0.0004664   2.953  0.00318 **
## prevmiYes    3.3117019  0.0558206  59.328  < 2e-16 ***
## prevstrkYes  0.1227498  0.0867712   1.415  0.15731
## prevhypYes   0.3953859  0.0325430  12.150  < 2e-16 ***
## hdlc        -0.0078128  0.0008175  -9.556  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5703 on 2290 degrees of freedom
## Multiple R-squared:  0.7732, Adjusted R-squared:  0.7717
## F-statistic: 520.4 on 15 and 2290 DF,  p-value: < 2.2e-16
```

2 Cross-Validation

```
require(statmod)
```

```
## Loading required package: statmod
```

```
## Warning: package 'statmod' was built under R version 3.6.3
```

```
# logitnorm_mean function
logitnorm_mean <- function(mu, sigma){
  v <- 1/(1+exp(-mu))
  alpha1 <- 1/((sigma^2)*(1-v))
  alpha2 <- 1/(v*(sigma^2))
  gqp <- gauss.quad.prob(n=10,dist="beta",alpha=alpha1,beta=alpha2)
  x <- gqp$nodes
  y <- gqp$weights
  g <- dnorm((log(x/(1-x))),mean=mu,sd=sigma,log = TRUE) - log(1-x) -
    dbeta(x, shape1=alpha1,shape2 = alpha2, log = TRUE)
```

```r
  sum(y*(exp(g)))
}


# compare Mstep to M2
M10 <- Mstep
M20 <- M2
Mnames <- expression(M[STEP], M[MLM])

# number of cross-validation replications
nreps <- 2e3
ntot <- nrow(fhs) # total number of observations
ntrain <- 1500 # for fitting MLE's
ntest <- ntot-ntrain # for out-of-sample prediction

# storage space
mspe1 <- rep(NA, nreps) # mspe for M10
mspe2 <- rep(NA, nreps) # mspe for M20
lambda1 <- rep(NA, nreps) # out-of-sample log-likelihood for M1
lambda2 <- rep(NA, nreps) # out-of-sample log-likelihood for M2

# cross-validation
system.time({
  for(ii in 1:nreps) {
    if(ii%%100 == 0) message("ii = ", ii)
    train.ind <- sample(ntot, ntrain) # training observations

    # fit the models on the subset of training data
    M10.cv <- update(M10, subset = train.ind)
    M20.cv <- update(M20, subset = train.ind)

    # out-of-sample log-likelihoods
    M10.sigma <- sqrt(sum(resid(M10.cv)^2)/ntrain) # MLE of sigma
    M20.sigma <- sqrt(sum(resid(M20.cv)^2)/ntrain)

    mu_1 <- predict(M10.cv, newdata = fhs[-train.ind,])
    mu_2 <- predict(M20.cv, newdata = fhs[-train.ind,])

    # out-of-sample residuals
    M10.res <- fhs$chdrisk[-train.ind] -
      sapply(1:ntest, function(ii) logitnorm_mean(mu_1[ii],M10.sigma))
    M20.res <- fhs$chdrisk[-train.ind] -
      sapply(1:ntest, function(ii) logitnorm_mean(mu_2[ii],M20.sigma))

    # mean-square prediction errors for each model
    mspe1[ii] <- mean(M10.res^2)
    mspe2[ii] <- mean(M20.res^2)

    # since res = y - pred, dnorm(y, pred, sd) = dnorm(res, 0, sd)
    lambda1[ii] <- sum(dnorm(M10.res, sd = M10.sigma, log = TRUE))
    lambda2[ii] <- sum(dnorm(M20.res, sd = M20.sigma, log = TRUE))
  }
})
```

```
## ii = 100
```

```
## ii = 200
```

```
## ii = 300
```

```
## ii = 400
```

```
## ii = 500
```

```
## ii = 600
```

```
## ii = 700
```

```
## ii = 800
```

```
## ii = 900
```

```
## ii = 1000
```

```
## ii = 1100
```

```
## ii = 1200
```

```
## ii = 1300
```

```
## ii = 1400
```

```
## ii = 1500
```

```
## ii = 1600
```

```
## ii = 1700
```

```
## ii = 1800
```

```
## ii = 1900
```

```
## ii = 2000
```

```
##    user  system elapsed
##  244.38    0.04  245.54
```