Week8-Exercises-Solutions

Exercise solutions

Use the tools above to investigate the ideal number of knots for the week 7 investigation between HDL and BMI.

Below shows the code and output for running a cubic spline with 3, 4, and 5 knots. We see that AIC and BIC consistently increase with increasing number of knots, and that 3 knots seems to generate the lowest AIC and BIC values. Therefore 3 knots is the most appropriate function form for BMI in this data.

Stata code and output

```
clear
use hersdata.dta

mkspline BMIsp3k = BMI, cubic nknots(3)
regress HDL BMIsp3k1 BMIsp3k2 age nonwhite smoking drinkany
estimates stats

mkspline BMIsp4k = BMI, cubic nknots(4)
regress HDL BMIsp4k1 BMIsp4k2 BMIsp4k3 age nonwhite smoking drinkany
estimates stats

mkspline BMIsp5k = BMI, cubic nknots(5)
regress HDL BMIsp5k1 BMIsp5k2 BMIsp5k3 BMIsp5k4 age nonwhite smoking drinkany
estimates stats
## . cl. use hersdata.dta
##
## .
## . mkspline BMIsp3k = BMI, cubic nknots(3)
##
## . regress HDL BMIsp3k1 BMIsp3k2 age nonwhite smoking drinkany
##
##       Source |       SS           df       MS      Number of obs   =     2,745
## -------------+----------------------------------   F(6, 2738)      =     40.46
##        Model |  38886.5157         6  6481.08595   Prob > F        =    0.0000
##     Residual |  438593.997     2,738  160.187727   R-squared       =    0.0814
```

```
## -------------+----------------------------------   Adj R-squared   =    0.0794
##        Total |  477480.512     2,744  174.008933   Root MSE        =    12.657
##
## ------------------------------------------------------------------------------
##          HDL | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##     BMIsp3k1 |  -.9349838   .1155016    -8.09   0.000    -1.161463   -.7085046
##     BMIsp3k2 |   .7100661   .1469379     4.83   0.000     .4219457    .9981864
##          age |   .1883881   .0380185     4.96   0.000     .1138403    .2629359
##     nonwhite |   2.477776   .7812025     3.17   0.002     .9459698    4.009581
##      smoking |  -2.089692   .7446461    -2.81   0.005    -3.549817   -.6295669
##     drinkany |   4.376112   .5039525     8.68   0.000     3.387946    5.364277
##        _cons |   60.64268   4.085344    14.84   0.000     52.63201    68.65334
## ------------------------------------------------------------------------------
##
## . estimates stats
##
## Akaike's information criterion and Bayesian information criterion
##
## -----------------------------------------------------------------------------
##        Model |          N   ll(null)  ll(model)      df        AIC        BIC
## -------------+---------------------------------------------------------------
##            . |      2,745  -10975.36  -10858.77       7   21731.53   21772.96
## -----------------------------------------------------------------------------
## Note: BIC uses N = number of observations. See [R] BIC note.
##
## .
## . mkspline BMIsp4k = BMI, cubic nknots(4)
##
## . regress HDL BMIsp4k1 BMIsp4k2 BMIsp4k3 age nonwhite smoking drinkany
##
##       Source |       SS           df       MS      Number of obs   =     2,745
## -------------+----------------------------------   F(7, 2737)      =     34.69
##        Model |  38911.9694         7  5558.85277   Prob > F        =    0.0000
##     Residual |  438568.543     2,737  160.236954   R-squared       =    0.0815
## -------------+----------------------------------   Adj R-squared   =    0.0791
##        Total |  477480.512     2,744  174.008933   Root MSE        =    12.658
##
## ------------------------------------------------------------------------------
##          HDL | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
```

```
##     BMIsp4k1 |  -1.024473   .2074958    -4.94   0.000    -1.431337   -.6176088
##     BMIsp4k2 |   1.279349   .8735961     1.46   0.143     -.433625    2.992324
##     BMIsp4k3 |   -2.02277    2.54743    -0.79   0.427    -7.01785    2.972311
##          age |   .1884047   .0380416     4.95   0.000     .1138116    .2629978
##     nonwhite |    2.46911   .7820285     3.16   0.002     .9356845    4.002536
##      smoking |  -2.097951   .7450014    -2.82   0.005    -3.558772   -.6371287
##     drinkany |   4.376638   .5040996     8.68   0.000     3.388184    5.365092
##        _cons |   62.58624   5.669147    11.04   0.000        51.47    73.70248
## ------------------------------------------------------------------------------
##
## . estimates stats
##
## Akaike's information criterion and Bayesian information criterion
##
## -----------------------------------------------------------------------------
##        Model |          N   ll(null)  ll(model)      df        AIC        BIC
## -------------+---------------------------------------------------------------
##            . |      2,745  -10975.36  -10858.69       8   21733.38   21780.72
## -----------------------------------------------------------------------------
## Note: BIC uses N = number of observations. See [R] BIC note.
##
## .
## . mkspline BMIsp5k = BMI, cubic nknots(5)
##
## . regress HDL BMIsp5k1 BMIsp5k2 BMIsp5k3 BMIsp5k4 age nonwhite smoking drinkany
##
##       Source |       SS           df       MS      Number of obs   =     2,745
## -------------+----------------------------------   F(8, 2736)      =     30.35
##        Model |  38913.5934         8  4864.19917   Prob > F        =    0.0000
##     Residual |  438566.919     2,736  160.294926   R-squared       =    0.0815
## -------------+----------------------------------   Adj R-squared   =    0.0788
##        Total |  477480.512     2,744  174.008933   Root MSE        =    12.661
##
## ------------------------------------------------------------------------------
##          HDL | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##     BMIsp5k1 |  -1.008258   .2823244    -3.57   0.000    -1.561849   -.4546676
##     BMIsp5k2 |   1.139488   2.424866     0.47   0.638    -3.615266    5.894242
##     BMIsp5k3 |  -.4761042   9.557886    -0.05   0.960    -19.21751     18.2653
##     BMIsp5k4 |  -1.757718   11.21143    -0.16   0.875    -23.74145    20.22601
##          age |   .1882574   .0380726     4.94   0.000     .1136035    .2629113
```

```
##     nonwhite |   2.469817   .7823079     3.16   0.002     .9358431    4.003791
##      smoking |  -2.097091   .7452066    -2.81   0.005    -3.558315   -.6358663
##     drinkany |   4.376239   .5041816     8.68   0.000     3.387624    5.364854
##        _cons |    62.2474   6.939817     8.97   0.000     48.63959    75.85521
## ------------------------------------------------------------------------------
##
## . estimates stats
##
## Akaike's information criterion and Bayesian information criterion
##
## ----------------------------------------------------------------------------
##        Model |          N   ll(null)  ll(model)      df        AIC        BIC
## -------------+--------------------------------------------------------------
##            . |      2,745  -10975.36  -10858.68       9   21735.37   21788.62
## ----------------------------------------------------------------------------
## Note: BIC uses N = number of observations. See [R] BIC note.
```

R code and output.

```
require(haven)
## Loading required package: haven
library(rms)
## Loading required package: Hmisc
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units

hers<-read_dta("https://www.dropbox.com/s/ndtd4o20qogq7fv/hersdata.dta?dl=1")
hers<-data.frame(hers)

# reduce the dataset and remove missing
hers1<-hers[,c("HDL", "BMI", "age", "nonwhite", "smoking", "drinkany")]
hers1<-na.omit(hers1)
dim(hers1)
## [1] 2745    6
# 2745 after removing the missing

ddist <- datadist(hers1)
options(datadist='ddist')
```

```r
# Run models with 3 - 7 knots
fit.3knots <-  lm(HDL ~ rcs(BMI,3) + age + nonwhite + smoking +drinkany, data = hers1)
fit.4knots <-  lm(HDL ~ rcs(BMI,4) + age + nonwhite + smoking +drinkany, data = hers1)
fit.5knots <-  lm(HDL ~ rcs(BMI,5) + age + nonwhite + smoking +drinkany, data = hers1)

AIC(fit.3knots)
## [1] 21733.53
AIC(fit.4knots)
## [1] 21735.38
AIC(fit.5knots)
## [1] 21737.37

BIC(fit.3knots)
## [1] 21780.87
BIC(fit.4knots)
## [1] 21788.63
BIC(fit.5knots)
## [1] 21796.54
```