Week7-Exercises-Solutions

Exercise solutions

Week 7

Investigation

Stata code and output

1) Standard regression analysis of *HDL* on *age*, *BMI*, *nonwhite*, *smoking* and *drinkany*. Note *age* and *BMI* have been centred in Table 4.20 so we do the same. A qudratic term in (centred) *BMI* so we will again follow Vittinghof et al.'s logic.

```
use hersdata.dta
summarize age, meanonly
gen agec = age - r(mean)
summarize BMI, meanonly
gen BMIc =  BMI- r(mean)
gen BMIc2 = BMIc^2
regress HDL BMIc BMIc2 agec nonwhite smoking drinkany
## . use hersdata.. summarize age, meanonly
##
## . gen agec = age - r(mean)
##
## . summarize BMI, meanonly
##
## . gen BMIc =  BMI- r(mean)
## (5 missing values generated)
##
## . gen BMIc2 = BMIc^2
## (5 missing values generated)
##
## . regress HDL BMIc BMIc2 agec nonwhite smoking drinkany
##
##       Source |       SS           df       MS      Number of obs   =     2,745
## -------------+----------------------------------   F(6, 2738)      =     39.99
##        Model |  38474.0926         6  6412.34877   Prob > F        =    0.0000
```

```
##     Residual |   439006.42     2,738  160.338356   R-squared       =    0.0806
## -------------+----------------------------------   Adj R-squared   =    0.0786
##        Total |  477480.512     2,744  174.008933   Root MSE        =    12.662
##
## ------------------------------------------------------------------------------
##          HDL | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##         BMIc |  -.5272427   .0507663   -10.39   0.000    -.6267868   -.4276985
##        BMIc2 |   .0242527   .0053231     4.56   0.000      .013815    .0346904
##         agec |   .1893209   .0380347     4.98   0.000     .1147414    .2639005
##     nonwhite |   2.494766   .7815733     3.19   0.001     .9622325    4.027299
##      smoking |  -2.070298   .7449086    -2.78   0.005    -3.530938   -.6096584
##     drinkany |   4.345096   .5041409     8.62   0.000     3.356561    5.333631
##        _cons |   47.80001   .3807692   125.54   0.000     47.05339    48.54663
## ------------------------------------------------------------------------------
```

we indeed observe that the quadratic term in centred BMI (called *BMIc2*) is significant suggesting that linearity might not be appropriate for BMI.

2) Fit a RCS in BMI with 4 knots while adjusting for the other covariates. Here BMI is not centred so we will keep it as is. We use *agec* as opposed to *age10* but feel free to scale age if you want.

```
clear
use hersdata.dta
summarize age, meanonly
gen agec = age - r(mean)
mkspline BMIsp = BMI, cubic nknots(4)
regress HDL BMIsp1 BMIsp2 BMIsp3  agec nonwhite smoking drinkany
test BMIsp2 BMIsp3
test BMIsp1 BMIsp2 BMIsp3
## . cl. use hersdata.dta
##
## . summarize age, meanonly
##
## . gen agec = age - r(mean)
##
## . mkspline BMIsp = BMI, cubic nknots(4)
##
## . regress HDL BMIsp1 BMIsp2 BMIsp3  agec nonwhite smoking drinkany
##
##       Source |       SS           df       MS      Number of obs   =     2,745
```

```
## -------------+----------------------------------   F(7, 2737)      =     34.69
##        Model |  38911.9694         7  5558.85278   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]
## -------------+----------------------------------------------------------------
##       BMIsp1 |  -1.024473   .2074958    -4.94   0.000    -1.431337   -.6176088
##       BMIsp2 |   1.279349   .8735961     1.46   0.143     -.433625    2.992324
##       BMIsp3 |   -2.02277    2.54743    -0.79   0.427     -7.01785    2.972311
##         agec |   .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 |   75.14314   4.854449    15.48   0.000     65.62439     84.6619
## ------------------------------------------------------------------------------
##
## . test BMIsp2 BMIsp3
##
##  ( 1)  BMIsp2 = 0
##  ( 2)  BMIsp3 = 0
##
##        F(  2,  2737) =   11.75
##             Prob > F =    0.0000
##
## . test BMIsp1 BMIsp2 BMIsp3
##
##  ( 1)  BMIsp1 = 0
##  ( 2)  BMIsp2 = 0
##  ( 3)  BMIsp3 = 0
##
##        F(  3,  2737) =   36.91
##             Prob > F =    0.0000
```

in Stata's syntax, *BMIsp1* is *age*. Although the additional spline terms (represented by *BMIsp2* and *BMIsp3* in the model) are not significant, the global test with 2 d.f. indicates that the splines are necessary $F = 11.75$, $p = 0.000$, highly significant result. The global effect of *BMI* is also highly significant and involves all the BMI-related terms (3 d.f), $F = 36.91$, $p = 0.000$. We let you reproduce the analysis presented in the book with 5 knots (although it's not requested).

3) Plot the fitted line for BMI with its 95% band.

Here you need to install *postrcspline* package to get a nice plot. Google helps with this, run the commad *ssc install postrcspline* (only once). Then use the command *mkspline2* is used to recreate the splines an plot them at specific values of the covariates.

```
clear
use hersdata.dta
summarize age, meanonly
gen agec = age - r(mean)
mkspline2 BMIsp = BMI, cubic nknots(4)
regress HDL BMIsp1 BMIsp2 BMIsp3  agec nonwhite smoking drinkany
adjustrcspline, at(agec=0 nonwhite=0 smoking=0 drinkany=0) title(Adjusted predictions)
## . cl. use hersdata.dta
##
## . summarize age, meanonly
##
## . gen agec = age - r(mean)
##
## . mkspline2 BMIsp = BMI, cubic nknots(4)
## command mkspline2 is unrecognized
## r(199);
##
## end of do-file
## r(199);
```

The command *adjustrcspline* produces the fitted line and its 95% CI. The spline is adjusted for the other covariates so you have to indicate values for those to get a plot. There is a default but it is probably better to choose your own values, e.g. we chose *agec=0* (i.e. *age=66.65*) and *nonwhite*, *smoking* and *drinkany* all set to 0. The plot, particularly the confidence band, will change if you choose other value but the shape remains the same.

4) Change the location of the 4 knots and refit the model. Try a different number of knots. Conclusion/interpretation.

We can fit a model with 4 knots placed differently (e.g. at age 18, 22, 25, 35)

```
clear
use hersdata.dta
summarize age, meanonly
gen agec = age - r(mean)
mkspline2 BMIsp = BMI, cubic knots(18 22 25 35)
regress HDL BMIsp1 BMIsp2 BMIsp3  agec nonwhite smoking drinkany
```

```
test BMIsp2 BMIsp3
adjustrcspline, at(agec=0 nonwhite=0 smoking=0 drinkany=0) title("4 knots at age 18, 22, 2
## . cl. use hersdata.dta
##
## . summarize age, meanonly
##
## . gen agec = age - r(mean)
##
## . mkspline2 BMIsp = BMI, cubic knots(18 22 25 35)
## command mkspline2 is unrecognized
## r(199);
##
## end of do-file
## r(199);
```

We can also fit a model with 5 knots placed at their default location.

```
clear
use hersdata.dta
summarize age, meanonly
gen agec = age - r(mean)
mkspline2 BMIsp = BMI, cubic nknots(5)
regress HDL BMIsp*  agec nonwhite smoking drinkany
test BMIsp2 BMIsp3 BMIsp4
adjustrcspline, at(agec=0 nonwhite=0 smoking=0 drinkany=0) title("5 knots with default loc
## . cl. use hersdata.dta
##
## . summarize age, meanonly
##
## . gen agec = age - r(mean)
##
## . mkspline2 BMIsp = BMI, cubic nknots(5)
## command mkspline2 is unrecognized
## r(199);
##
## end of do-file
## r(199);
```

Once again, a spline is necessary. We do not have the tools yet to discriminate between the different fitted splines but they look rather similar. The data supports the use of splines. HDL decreases markedly with BMI until the age of 30-35 where the decrease is not as steep. This is after adjustment for (centred) *age*, *nonwhite*, *smoking* and *drinkany*.

5) Do we need a RCS model for age?

```
clear
use hersdata.dta
mkspline2 BMIsp = BMI, cubic nknots(4)
mkspline2 agesp = age, cubic nknots(4)
regress HDL BMIsp* agesp*  nonwhite smoking drinkany
test agesp2 agesp3
adjustrcspline, at(BMI= 27.75 nonwhite=0 smoking=0 drinkany=0) title("4 knots with default
## . cl. use hersdata.dta
##
## . mkspline2 BMIsp = BMI, cubic nknots(4)
## command mkspline2 is unrecognized
## r(199);
##
## end of do-file
## r(199);
```

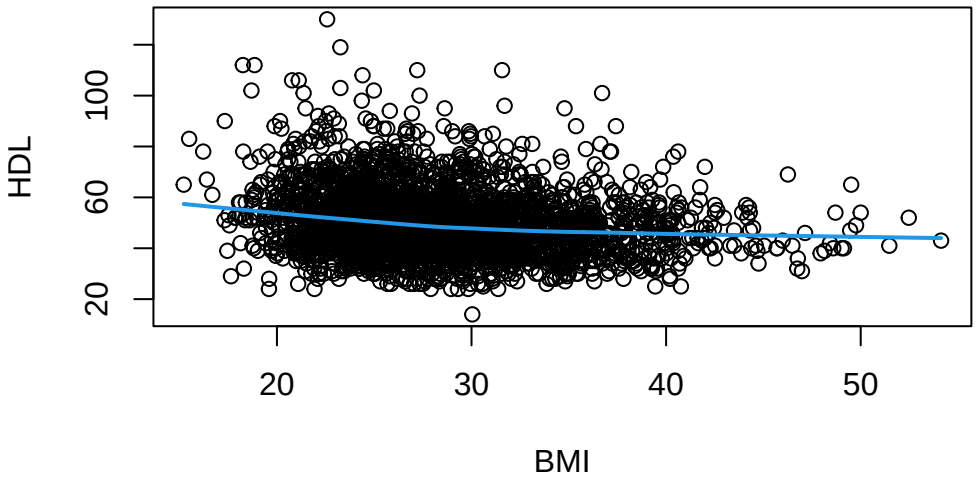
We keep the standard RCS(4) model for BMI and added a spline in age (i.e RCS(4) in age). There is no evidence in the data that the additional terms in age are needed. The command *test agesp2 agesp3* returns a F-test of 0.69, p=0.499 and the spline in age (plot omitted here) is rather straight. We would keep *age* alone in the model but keep the RCS(4) in *BMI* in the final model.

R code and output

1) Standard regression analysis of *HDL* on *age*, *BMI*, *nonwhite*, *smoking* and *drinkany*. Note *age* and *BMI* have been centred in Table 4.20 so we do the same. A quadratic term in (centred) *BMI* so we will again follow Vittinghof et al.'s logic.

```
require(haven)
## Loading required package: haven
hers<-read_dta("hersdata.dta")
hers<-data.frame(hers)

hers$agec<-hers$age-mean(hers$age,na.rm=TRUE)
hers$BMIc<-hers$BMI-mean(hers$BMI,na.rm=TRUE)
hers$BMIc2<-hers$BMIc^2

# reduce the dataset and remove missing
hers1<-cbind(hers$HDL, hers$BMI, hers$age, hers$BMIc, hers$BMIc2, hers$agec, hers$nonwhite
hers1<-data.frame(hers1)
colnames(hers1)=c("HDL", "BMI", "age", "BMIc", "BMIc2", "agec", "nonwhite", "smoking", "dr
```

```r
hers1<-na.omit(hers1)
dim(hers1)
## [1] 2745    9
# 2745 after removing the missing

# exploratory analysis
plot(HDL ~ BMI, data=hers1)
lines(lowess(hers1$HDL ~ hers1$BMI), col=4,lwd=2)
```




```r
# fit quadratic model in BMI
# directly
fit.quad <- lm(HDL ~ BMIc + BMIc2 + agec + nonwhite + smoking + drinkany, data = hers1)
summary(fit.quad)
##
## Call:
## lm(formula = HDL ~ BMIc + BMIc2 + agec + nonwhite + smoking +
##     drinkany, data = hers1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.852  -8.694  -1.645   6.674  81.872
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.800010   0.380769 125.535  < 2e-16 ***
## BMIc        -0.527243   0.050766 -10.386  < 2e-16 ***
## BMIc2        0.024253   0.005323   4.556 5.44e-06 ***
## agec         0.189321   0.038035   4.978 6.84e-07 ***
## nonwhite     2.494766   0.781573   3.192  0.00143 **
## smoking     -2.070298   0.744909  -2.779  0.00549 **
## drinkany     4.345096   0.504141   8.619  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.66 on 2738 degrees of freedom
## Multiple R-squared:  0.08058,    Adjusted R-squared:  0.07856
## F-statistic: 39.99 on 6 and 2738 DF,  p-value: < 2.2e-16
# using poly() that creates the quadratic term
fit.quad <- lm(HDL ~ poly(BMIc,2, raw="TRUE") + agec + nonwhite + smoking + drinkany, data
summary(fit.quad) # same
##
## Call:
## lm(formula = HDL ~ poly(BMIc, 2, raw = "TRUE") + agec + nonwhite +
##     smoking + drinkany, data = hers1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.852  -8.694  -1.645   6.674  81.872
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    47.800010   0.380769 125.535  < 2e-16 ***
## poly(BMIc, 2, raw = "TRUE")1 -0.527243   0.050766 -10.386  < 2e-16 ***
## poly(BMIc, 2, raw = "TRUE")2  0.024253   0.005323   4.556 5.44e-06 ***
## agec                           0.189321   0.038035   4.978 6.84e-07 ***
## nonwhite                       2.494766   0.781573   3.192  0.00143 **
## smoking                       -2.070298   0.744909  -2.779  0.00549 **
## drinkany                       4.345096   0.504141   8.619  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.66 on 2738 degrees of freedom
```

```
## Multiple R-squared:  0.08058,    Adjusted R-squared:  0.07856
## F-statistic: 39.99 on 6 and 2738 DF,  p-value: < 2.2e-16
anova(fit.quad)
## Analysis of Variance Table
##
## Response: HDL
##                               Df Sum Sq Mean Sq F value    Pr(>F)
## poly(BMIc, 2, raw = "TRUE")    2  20847 10423.7 65.0105 < 2.2e-16 ***
## agec                           1   3802  3802.2 23.7138 1.181e-06 ***
## nonwhite                       1    575   574.5  3.5831  0.058477 .
## smoking                        1   1339  1339.4  8.3538  0.003879 **
## drinkany                       1  11911 11910.6 74.2839 < 2.2e-16 ***
## Residuals                   2738 439006   160.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We indeed observe that the quadratic term in centred BMI (called *BMIc2* in the first fit) is significant suggesting that linearity might not be appropriate for BMI.

2) Fit a RCS in BMI with 4 knots while adjusting for the other covariates. Here BMI is not centred so we will keep it as is. We use *agec* as opposed to *age10* but feel free to rescale age if you want.

```
library(rms)
## Loading required package: Hmisc
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
ddist <- datadist(hers1)
options(datadist='ddist')
fit1 <-  ols(HDL ~ rcs(BMI,4) + agec + nonwhite + smoking +drinkany, data = hers1)
fit1
## Linear Regression Model
##
## ols(formula = HDL ~ rcs(BMI, 4) + agec + nonwhite + smoking +
##     drinkany, data = hers1)
##
##                  Model Likelihood    Discrimination
##                        Ratio Test           Indexes
## Obs     2745    LR chi2    233.34    R2       0.081
```

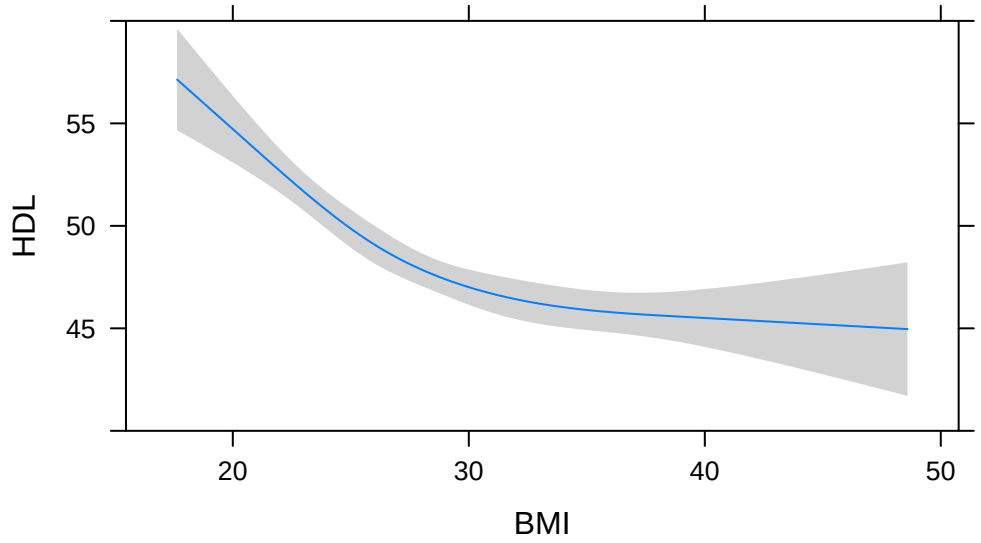
```
## sigma12.6585    d.f.            7    R2 adj   0.079
## d.f.    2737    Pr(> chi2) 0.0000    g        4.253
##
## Residuals
##
##     Min      1Q  Median      3Q     Max
## -34.691  -8.698  -1.600   6.669  81.699
##
##
##           Coef    S.E.   t     Pr(>|t|)
## Intercept 75.1426 4.8576 15.47 <0.0001
## BMI       -1.0244 0.2076 -4.93 <0.0001
## BMI'       1.2791 0.8755  1.46 0.1441
## BMI''     -2.0167 2.5455 -0.79 0.4283
## agec       0.1884 0.0380  4.95 <0.0001
## nonwhite   2.4691 0.7820  3.16 0.0016
## smoking   -2.0979 0.7450 -2.82 0.0049
## drinkany   4.3766 0.5041  8.68 <0.0001
anova(fit1)
##                 Analysis of Variance          Response: HDL
##
##  Factor     d.f. Partial SS MS        F     P
##  BMI           3  17741.225  5913.742 36.91 <.0001
##   Nonlinear    2   3766.168  1883.084 11.75 <.0001
##  agec          1   3930.360  3930.360 24.53 <.0001
##  nonwhite      1   1597.339  1597.339  9.97 0.0016
##  smoking       1   1270.680  1270.680  7.93 0.0049
##  drinkany      1  12078.483 12078.483 75.38 <.0001
##  REGRESSION    7  38911.929  5558.847 34.69 <.0001
##  ERROR      2737 438568.583   160.237
```

Although the additional terms (represented by *BMI'* and *BMI''* in the model) are not significant, the global test with 2 d.f. provided by the *anova* command indicates that the additional terms are necessary $F = 11.75$, $p = 0.000$, a highly significant result. This is typical of spline-based analysis. We let you reproduce the analysis presented in the book with 5 knots (although it's not requested).

3) Plot the fitted line for BMI with its 95% band.

This is straightforward in R and the command fixes the other covariates at default values. The median of each predictor other than BMI is typically used.

```
plot(Predict(fit1, BMI))
```

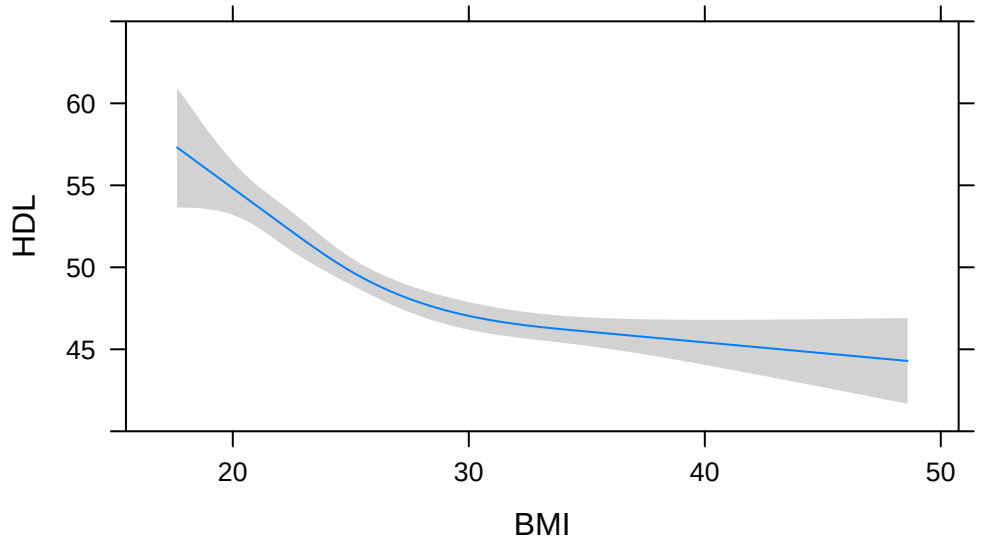


4) Change the location of the 4 knots and refit the model. Try a different number of knots. Conclusion/interpretation.

We can fit a model with 4 knots placed differently (e.g. at age 18, 22, 25, 35) or with 5 knots (default location) as examples.

```
# 4 different knots at 18, 22, 25, 35
fit2 <-  ols(HDL ~ rcs(BMI,c(18,22,25,35)) + agec + nonwhite + smoking +drinkany, data = h
fit2
## Linear Regression Model
##
## ols(formula = HDL ~ rcs(BMI, c(18, 22, 25, 35)) + agec + nonwhite +
##     smoking + drinkany, data = hers1)
##
##                  Model Likelihood    Discrimination
##                        Ratio Test           Indexes
## Obs     2745    LR chi2    233.01    R2       0.081
## sigma12.6592    d.f.            7    R2 adj   0.079
## d.f.    2737    Pr(> chi2) 0.0000    g        4.243
##
```

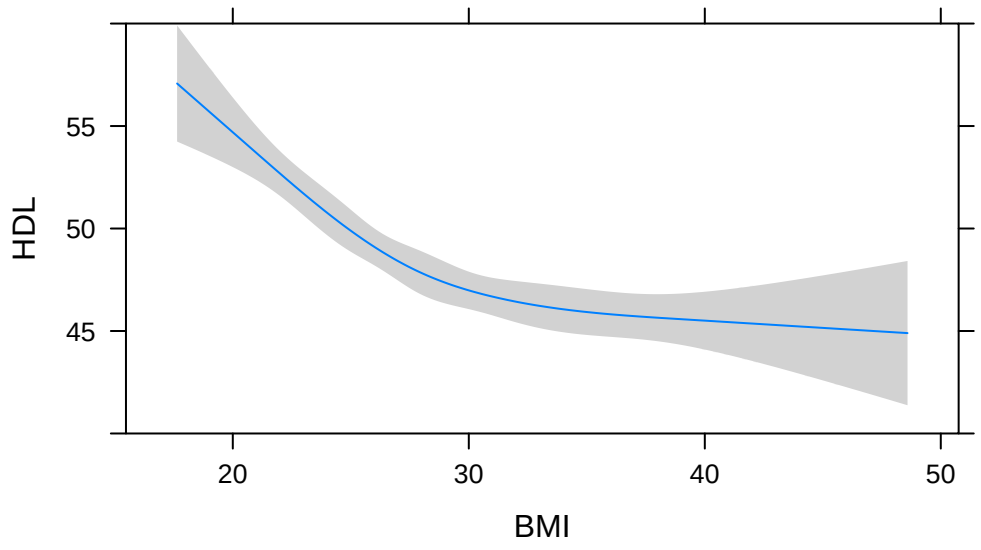
```
## Residuals
##
##     Min      1Q  Median      3Q     Max
## -34.715  -8.701  -1.598   6.689  81.688
##
##
##           Coef    S.E.    t     Pr(>|t|)
## Intercept 75.8527 11.6628  6.50 <0.0001
## BMI       -1.0549  0.5620 -1.88 0.0606
## BMI'       0.0059  2.4600  0.00 0.9981
## BMI''      2.2622  6.2588  0.36 0.7178
## agec       0.1878  0.0380  4.94 <0.0001
## nonwhite   2.4542  0.7825  3.14 0.0017
## smoking   -2.0996  0.7451 -2.82 0.0049
## drinkany   4.3813  0.5041  8.69 <0.0001
anova(fit2)
##                 Analysis of Variance          Response: HDL
##
##  Factor     d.f. Partial SS MS         F     P
##  BMI           3  17687.916  5895.9721 36.79 <.0001
##   Nonlinear    2   3712.859  1856.4295 11.58 <.0001
##  agec          1   3906.439  3906.4395 24.38 <.0001
##  nonwhite      1   1576.466  1576.4655  9.84 0.0017
##  smoking       1   1272.520  1272.5205  7.94 0.0049
##  drinkany      1  12105.088 12105.0879 75.54 <.0001
##  REGRESSION    7  38858.620  5551.2315 34.64 <.0001
##  ERROR      2737 438621.892   160.2564
plot(Predict(fit2, BMI))
```



Adjusted to:agec=0.3514 nonwhite=0 smoking=0 drinkany=0

```
# 5  knots, default location
fit3 <-  ols(HDL ~ rcs(BMI,5) + agec + nonwhite + smoking +drinkany, data = hers1)
fit3
## Linear Regression Model
##
## ols(formula = HDL ~ rcs(BMI, 5) + agec + nonwhite + smoking +
##     drinkany, data = hers1)
##
##                  Model Likelihood    Discrimination
##                        Ratio Test           Indexes
## Obs     2745    LR chi2    233.35    R2       0.081
## sigma12.6608    d.f.            8    R2 adj   0.079
## d.f.    2736    Pr(> chi2) 0.0000    g        4.254
##
## Residuals
##
##     Min      1Q  Median      3Q     Max
## -34.663  -8.708  -1.602   6.681  81.681
##
##
##           Coef    S.E.    t     Pr(>|t|)
```

```
## Intercept 74.7924  6.3468 11.78 <0.0001
## BMI       -1.0082  0.2826 -3.57 0.0004
## BMI'       1.1378  2.4348  0.47 0.6403
## BMI''     -0.4679  9.5871 -0.05 0.9611
## BMI'''    -1.7592 11.2122 -0.16 0.8753
## agec       0.1883  0.0381  4.94 <0.0001
## nonwhite   2.4698  0.7823  3.16 0.0016
## smoking   -2.0971  0.7452 -2.81 0.0049
## drinkany   4.3762  0.5042  8.68 <0.0001
anova(fit3)
##                 Analysis of Variance          Response: HDL
##
##  Factor     d.f. Partial SS MS         F     P
##  BMI           4  17742.871  4435.7176 27.67 <.0001
##   Nonlinear    3   3767.813  1255.9377  7.84 <.0001
##  agec          1   3919.250  3919.2496 24.45 <.0001
##  nonwhite      1   1597.700  1597.7004  9.97 0.0016
##  smoking       1   1269.392  1269.3923  7.92 0.0049
##  drinkany      1  12076.748 12076.7481 75.34 <.0001
##  REGRESSION    8  38913.574  4864.1968 30.35 <.0001
##  ERROR      2736 438566.938   160.2949
plot(Predict(fit3, BMI))
```



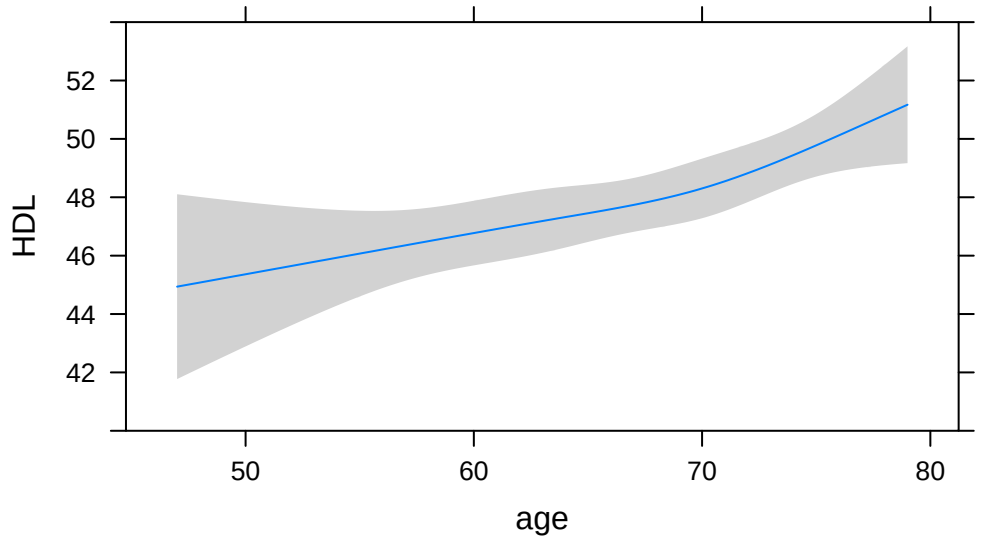
Adjusted to:agec=0.3514 nonwhite=0 smoking=0 drinkany=0

Once again, a spline is necessary irrespective of the number of knots or their location. We do not have the tools yet to discriminate between the different spline fits but they look rather similar. The data supports the use of splines. HDL decreases markedly with BMI until the age of 30-35 where the decrease is not as steep. This is after adjustment for (centred) *age*, *nonwhite*, *smoking* and *drinkany*.

5) Do we need a RCS model for age?

```
fit4 <-  ols(HDL ~ rcs(BMI,4) + rcs(age,4) + nonwhite + smoking + drinkany, data = hers1)
fit4
## Linear Regression Model
##
## ols(formula = HDL ~ rcs(BMI, 4) + rcs(age, 4) + nonwhite + smoking +
##     drinkany, data = hers1)
##
##                  Model Likelihood    Discrimination
##                        Ratio Test           Indexes
## Obs     2745    LR chi2    234.74    R2       0.082
## sigma12.6599    d.f.            9    R2 adj   0.079
## d.f.    2735    Pr(> chi2) 0.0000    g        4.261
##
## Residuals
##
##     Min      1Q  Median      3Q     Max
## -35.126  -8.725  -1.568   6.713  81.549
##
##
##            Coef    S.E.    t     Pr(>|t|)
## Intercept 65.4523 9.1246  7.17 <0.0001
## BMI       -1.0234 0.2077 -4.93 <0.0001
## BMI'       1.2914 0.8757  1.47 0.1404
## BMI''     -2.0651 2.5461 -0.81 0.4174
## age        0.1416 0.1291  1.10 0.2731
## age'      -0.0163 0.3297 -0.05 0.9604
## age''      0.4996 1.4114  0.35 0.7234
## nonwhite   2.4439 0.7825  3.12 0.0018
## smoking   -2.1206 0.7462 -2.84 0.0045
## drinkany   4.3671 0.5043  8.66 <0.0001
anova(fit4)
##                 Analysis of Variance          Response: HDL
##
##  Factor          d.f. Partial SS MS         F     P
```

```
##  BMI                3  17600.372  5866.7908 36.61 <.0001
##   Nonlinear         2   3716.043  1858.0214 11.59 <.0001
##  age                3   4153.049  1384.3495  8.64 <.0001
##   Nonlinear         2    222.689   111.3445  0.69 0.4993
##  nonwhite           1   1563.379  1563.3792  9.75 0.0018
##  smoking            1   1294.234  1294.2337  8.08 0.0045
##  drinkany           1  12016.880 12016.8800 74.98 <.0001
##  TOTAL NONLINEAR    4   3988.857   997.2142  6.22 0.0001
##  REGRESSION         9  39134.618  4348.2909 27.13 <.0001
##  ERROR           2735 438345.894   160.2727
plot(Predict(fit4, age))
```



Adjusted to:BMI=27.74 nonwhite=0 smoking=0 drinkany=0

We keep the standard RCS(4) model for BMI and added an age spline (i.e RCS(4) in age). There is no evidence in the data that the spline in age is needed. The *anova* command returns a F-test of 0.69 for the non-linear component in age, p=0.499. This is also confirmed when we look at the plot, the effect of age appearing linear. We would keep *age* alone in the model but keep the RCS(4) in *BMI* in the final model.

Bootstrap investigation

Part A) of the investigation is mainly running the code to get familiar with bootstrapping in the regression. The remainder should be straightforward commands.

Stata code and output

1) standard regression of *HDL* on *BMI*, *age* and *drinkany* after removing diabetic patients. Examination of the residuals.

```
use hersdata.dta
drop if diabetes ==1
drop if mi(HDL) | mi(BMI) | mi(age) | mi(drinkany)
keep HDL BMI age drinkany
regress HDL BMI age drinkany
predict res, res
qnorm(res)
## . use hersdata.. drop if diabetes ==1
## (731 observations deleted)
##
## . drop if mi(HDL) | mi(BMI) | mi(age) | mi(drinkany)
## (11 observations deleted)
##
## . keep HDL BMI age drinkany
##
## . regress HDL BMI age drinkany
##
##       Source |       SS           df       MS      Number of obs   =     2,021
## -------------+----------------------------------   F(3, 2017)      =     46.91
##        Model |  24006.0684         3  8002.02279   Prob > F        =    0.0000
##     Residual |  344071.218     2,017  170.585631   R-squared       =    0.0652
## -------------+----------------------------------   Adj R-squared   =    0.0638
##        Total |  368077.286     2,020  182.216478   Root MSE        =    13.061
##
## ------------------------------------------------------------------------------
##          HDL | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.4036859   .0571063    -7.07   0.000    -.5156793   -.2916924
##          age |   .2086808   .0437416     4.77   0.000     .1228974    .2944643
##     drinkany |   4.502504   .5880671     7.66   0.000     3.349222    5.655787
##        _cons |   46.68225   3.571831    13.07   0.000     39.67739    53.68712
## ------------------------------------------------------------------------------
##
## . predict res, res
##
## . qnorm(res)
```

We clearly seem some upward curvature in the residuals (plot omitted here). Given the large

sample size, normality is not that critical but we are going to check that inference is indeed valid using the bootstrap.

2) Read Part A) of the code, run it and draw a histogram of the bootstrap samples (or replicates) for each of the coefficients

```
use hersdata.dta
drop if diabetes ==1
drop if mi(HDL) | mi(BMI) | mi(age) | mi(drinkany)
keep HDL BMI age drinkany
regress HDL BMI age drinkany
matrix observe= (_b[_cons], _b[BMI], _b[age], _b[drinkany])
matrix list observe
capture program drop myboot2
program define myboot2, rclass
 preserve
  bsample
    regress HDL BMI age drinkany
      return scalar b0 = _b[_cons]
    return scalar b1 = _b[BMI]
    return scalar b2 = _b[age]
    return scalar b3 = _b[drinkany]
 restore
end
** simulation = resampling the data using the program myboot2
simulate  b0=r(b0) b1=r(b1) b2=r(b2) b3=r(b3), reps(1000) seed(12345): myboot2
desc
hist b0
qnorm(b0)
hist b1
qnorm(b1)
hist b2
qnorm(b2)
hist b3
qnorm(b3)
bstat, stat(observe) n(2021)
estat bootstrap, percentile
estat bootstrap, all
## . use hersdata.. drop if diabetes ==1
## (731 observations deleted)
##
## . drop if mi(HDL) | mi(BMI) | mi(age) | mi(drinkany)
```

```
## (11 observations deleted)
##
## . keep HDL BMI age drinkany
##
## . regress HDL BMI age drinkany
##
##       Source |       SS           df       MS      Number of obs   =     2,021
## -------------+----------------------------------   F(3, 2017)      =     46.91
##        Model |  24006.0684         3  8002.02279   Prob > F        =    0.0000
##     Residual |  344071.218     2,017  170.585631   R-squared       =    0.0652
## -------------+----------------------------------   Adj R-squared   =    0.0638
##        Total |  368077.286     2,020  182.216478   Root MSE        =    13.061
##
## ------------------------------------------------------------------------------
##          HDL | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.4036859   .0571063    -7.07   0.000    -.5156793   -.2916924
##          age |   .2086808   .0437416     4.77   0.000     .1228974    .2944643
##     drinkany |   4.502504   .5880671     7.66   0.000     3.349222    5.655787
##        _cons |   46.68225   3.571831    13.07   0.000     39.67739    53.68712
## ------------------------------------------------------------------------------
##
## . matrix observe= (_b[_cons], _b[BMI], _b[age], _b[drinkany])
##
## . matrix list observe
##
## observe[1,4]
##             c1          c2          c3          c4
## r1   46.682253  -.40368587   .20868084   4.5025044
##
## . capture program drop myboot2
##
## . program define myboot2, rclass
##   1.  preserve
##   2.   bsample
##   3.     regress HDL BMI age drinkany
##   4.           return scalar b0 = _b[_cons]
##   5.     return scalar b1 = _b[BMI]
##   6.     return scalar b2 = _b[age]
##   7.         return scalar b3 = _b[drinkany]
##   8.  restore
```

```
##   9. end
##
## . ** simulation = resampling the data using the program myboot2
## . simulate  b0=r(b0) b1=r(b1) b2=r(b2) b3=r(b3), reps(1000) seed(12345): myboot2
##
##       Command: myboot2
##            b0: r(b0)
##            b1: r(b1)
##            b2: r(b2)
##            b3: r(b3)
##
## Simulations (1,000)
## ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
## ..................................................    50
## ..................................................   100
## ..................................................   150
## ..................................................   200
## ..................................................   250
## ..................................................   300
## ..................................................   350
## ..................................................   400
## ..................................................   450
## ..................................................   500
## ..................................................   550
## ..................................................   600
## ..................................................   650
## ..................................................   700
## ..................................................   750
## ..................................................   800
## ..................................................   850
## ..................................................   900
## ..................................................   950
## .................................................. 1,000
##
## . desc
##
## Contains data
##  Observations:         1,000                  simulate: myboot2
##     Variables:             4                  22 Apr 2023 14:32
## ------------------------------------------------------------------------------
## Variable      Storage   Display    Value
```

```
##     name         type    format    label      Variable label
## ------------------------------------------------------------------------------
## b0              float   %9.0g                 r(b0)
## b1              float   %9.0g                 r(b1)
## b2              float   %9.0g                 r(b2)
## b3              float   %9.0g                 r(b3)
## ------------------------------------------------------------------------------
## Sorted by:
##
## . hist b0
## (bin=29, start=36.100193, width=.76592847)
##
## . qnorm(b0)
##
## . hist b1
## (bin=29, start=-.58022505, width=.01300306)
##
## . qnorm(b1)
##
## . hist b2
## (bin=29, start=.07178009, width=.00905334)
##
## . qnorm(b2)
##
## . hist b3
## (bin=29, start=2.7042031, width=.12309739)
##
## . qnorm(b3)
##
## . bstat, stat(observe) n(2021)
##
## Bootstrap results                                       Number of obs = 2,021
##                                                         Replications  = 1,000
##
## ------------------------------------------------------------------------------
##              |   Observed   Bootstrap                         Normal-based
##              | coefficient  std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##           b0 |   46.68225   3.455292    13.51   0.000     39.91001     53.4545
##           b1 |  -.4036859   .0548886    -7.35   0.000    -.5112656   -.2961062
##           b2 |   .2086808   .0430616     4.85   0.000     .1242817      .29308
```

```
##           b3 |   4.502504   .5957828     7.56   0.000     3.334792    5.670217
## ------------------------------------------------------------------------------
##
## . estat bootstrap, percentile
##
## Bootstrap results                                 Number of obs     =      2,021
##                                                   Replications      =       1000
##
## ------------------------------------------------------------------------------
##              |    Observed               Bootstrap
##              | coefficient       Bias    std. err.  [95% conf. interval]
## -------------+----------------------------------------------------------------
##           b0 |   46.682253   .1759177   3.4552918    40.42489   54.07221   (P)
##           b1 |  -.40368587  -.0038854    .0548886   -.5188768  -.2992493   (P)
##           b2 |   .20868084  -.0009771   .04306157     .117856   .2910675   (P)
##           b3 |   4.5025044   .0024019   .59578279    3.368253   5.692966   (P)
## ------------------------------------------------------------------------------
## Key: P: Percentile
##
## . estat bootstrap, all
##
## Bootstrap results                                 Number of obs     =      2,021
##                                                   Replications      =       1000
##
## ------------------------------------------------------------------------------
##              |    Observed               Bootstrap
##              | coefficient       Bias    std. err.  [95% conf. interval]
## -------------+----------------------------------------------------------------
##           b0 |   46.682253   .1759177   3.4552918    39.91001    53.4545   (N)
##              |                                       40.42489   54.07221   (P)
##              |                                       39.84275   53.29212  (BC)
##           b1 |  -.40368587  -.0038854    .0548886   -.5112656  -.2961062   (N)
##              |                                      -.5188768  -.2992493   (P)
##              |                                       -.507547  -.2924751  (BC)
##           b2 |   .20868084  -.0009771   .04306157    .1242817     .29308   (N)
##              |                                        .117856   .2910675   (P)
##              |                                       .1173203   .2900742  (BC)
##           b3 |   4.5025044   .0024019   .59578279    3.334792   5.670217   (N)
##              |                                       3.368253   5.692966   (P)
##              |                                       3.405528   5.748656  (BC)
## ------------------------------------------------------------------------------
```

```
## Key:  N: Normal
##       P: Percentile
##      BC: Bias-corrected
```

The dataset contains R=1000 replicates or bootstrap samples (1000x4 since we have 4 coefficients), see the column names b0, b1, b2 and b3 as described by *desc*. Of course, you can change the number of replicates (e.g. use *reps(3000)* if you want 3000 replicates. R should be chosen large enough (at least 1000) for 95% confidence intervals. Histograms and normality plots follow easily. The plots have been omitted but you can check that the histograms are fairly symmetric and the distributions appear to be be normal. The commands *btsat* and *estat* give you the various CIs depending on the option you choose. When using *bstat* you need to specify the number of observations via the option *n(2021)*. To be updated with a different dataset.

3) direct calculation of the percentile 95% CI using a one-line command.

We can calculate the 2.5% and 97.5% centile for each variable (b0, b1, b2, b3) to get the required percentile CIs. The summary provided by Stata is also provided.

```
clear
use all_replicates
** these two lines are not needed when you run the code directly
** all replicates are saved in all_replicates.dta to simplify writing here
centile b0, centile(2.5 97.5)
centile b1, centile(2.5 97.5)
centile b2, centile(2.5 97.5)
centile b3, centile(2.5 97.5)
## . cl. use all_replicates
## (simulate: myboot2)
##
## . ** these two lines are not needed when you run the code directly
## . ** all replicates are saved in all_replicates.dta to simplify writing here
## . centile b0, centile(2.5 97.5)
##
##                                                       Binom. interp.
##     Variable |       Obs  Percentile    Centile        [95% conf. interval]
## -------------+-------------------------------------------------------------
##           b0 |     1,000        2.5    40.41275        39.61048    40.73114
##              |                 97.5    54.10855        53.35608    54.85093
##
## . centile b1, centile(2.5 97.5)
##
```

```
##                                                              Binom. interp.
##     Variable |       Obs  Percentile    Centile        [95% conf. interval]
## -------------+-------------------------------------------------------------
##           b1 |     1,000        2.5    -.519255       -.5326652   -.5059432
##              |                 97.5   -.2989461       -.3045337   -.2892334
##
## . centile b2, centile(2.5 97.5)
##
##                                                              Binom. interp.
##     Variable |       Obs  Percentile    Centile        [95% conf. interval]
## -------------+-------------------------------------------------------------
##           b2 |     1,000        2.5    .1173471        .1105016    .1268478
##              |                 97.5     .292011        .2845542    .3025858
##
## . centile b3, centile(2.5 97.5)
##
##                                                              Binom. interp.
##     Variable |       Obs  Percentile    Centile        [95% conf. interval]
## -------------+-------------------------------------------------------------
##           b3 |     1,000        2.5    3.366034        3.309025     3.42384
##              |                 97.5    5.695031        5.602746    5.822673
```

The results are very similar to Stata's (tiny differences), e.g. the percentile 95% CI for BMI is (-.519 ; -.299).

4) Using the Stata built-in command to avoid having to do the resampling "by hand'". Use Part B) and compare to the standard analysis. You can decide to display all 3 types or only one by speciying the *bootstrap* or *estat* command as follows:

```
clear
use hersdata.dta
drop if diabetes ==1
drop if mi(HDL) | mi(BMI) | mi(age) | mi(drinkany)
keep HDL BMI age drinkany
** display the 3 types (with R=1000 replicates)
bootstrap, reps(1000)  seed(12345): regress HDL BMI age drinkany
estat bootstrap, all
**
** only normal (R=1000)
bootstrap, reps(1000)  seed(12345): regress HDL BMI age drinkany
estat bootstrap, normal
** only percentile (R=1000)
```

```
bootstrap, reps(1000)  seed(12345): regress HDL BMI age drinkany
estat bootstrap, percentile
** only BCa (R=1000), slightly DIFFERENT command
bootstrap, bca reps(1000)  seed(12345): regress HDL BMI age drinkany
estat bootstrap
## . cl. use hersdata.dta
##
## . drop if diabetes ==1
## (731 observations deleted)
##
## . drop if mi(HDL) | mi(BMI) | mi(age) | mi(drinkany)
## (11 observations deleted)
##
## . keep HDL BMI age drinkany
##
## . ** display the 3 types (with R=1000 replicates)
## . bootstrap, reps(1000)  seed(12345): regress HDL BMI age drinkany
## (running regress on estimation sample)
##
## Bootstrap replications (1,000)
## ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
## ..................................................    50
## ..................................................   100
## ..................................................   150
## ..................................................   200
## ..................................................   250
## ..................................................   300
## ..................................................   350
## ..................................................   400
## ..................................................   450
## ..................................................   500
## ..................................................   550
## ..................................................   600
## ..................................................   650
## ..................................................   700
## ..................................................   750
## ..................................................   800
## ..................................................   850
## ..................................................   900
## ..................................................   950
## .................................................. 1,000
```

```
##
## Linear regression                                       Number of obs =   2,021
##                                                         Replications  =   1,000
##                                                         Wald chi2(3)  =  122.77
##                                                         Prob > chi2   =  0.0000
##                                                         R-squared     =  0.0652
##                                                         Adj R-squared =  0.0638
##                                                         Root MSE      = 13.0608
##
## ------------------------------------------------------------------------------
##              |   Observed   Bootstrap                         Normal-based
##          HDL | coefficient  std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.4036859   .0548886    -7.35   0.000    -.5112656   -.2961062
##          age |   .2086808   .0430616     4.85   0.000     .1242817      .29308
##     drinkany |   4.502504   .5957828     7.56   0.000     3.334792    5.670217
##        _cons |   46.68225   3.455292    13.51   0.000     39.91001     53.4545
## ------------------------------------------------------------------------------
##
## . estat bootstrap, all
##
## Linear regression                               Number of obs     =      2,021
##                                                 Replications      =       1000
##
## ------------------------------------------------------------------------------
##              |    Observed               Bootstrap
##          HDL | coefficient       Bias    std. err.  [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.40368587  -.0038854    .0548886   -.5112656  -.2961062   (N)
##              |                                      -.5188768  -.2992493   (P)
##              |                                       -.507547  -.2924751  (BC)
##          age |   .20868084  -.0009771   .04306157    .1242817     .29308   (N)
##              |                                        .117856   .2910675   (P)
##              |                                       .1173203   .2900742  (BC)
##     drinkany |   4.5025044   .0024019   .59578279    3.334792   5.670217   (N)
##              |                                       3.368253   5.692966   (P)
##              |                                       3.405528   5.748656  (BC)
##        _cons |   46.682253   .1759177   3.4552918    39.91001    53.4545   (N)
##              |                                       40.42489   54.07221   (P)
##              |                                       39.84275   53.29212  (BC)
## ------------------------------------------------------------------------------
```

```
## Key:  N: Normal
##       P: Percentile
##      BC: Bias-corrected
##
## . **
## . ** only normal (R=1000)
## . bootstrap, reps(1000)  seed(12345): regress HDL BMI age drinkany
## (running regress on estimation sample)
##
## Bootstrap replications (1,000)
## ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
## ..................................................    50
## ..................................................   100
## ..................................................   150
## ..................................................   200
## ..................................................   250
## ..................................................   300
## ..................................................   350
## ..................................................   400
## ..................................................   450
## ..................................................   500
## ..................................................   550
## ..................................................   600
## ..................................................   650
## ..................................................   700
## ..................................................   750
## ..................................................   800
## ..................................................   850
## ..................................................   900
## ..................................................   950
## .................................................. 1,000
##
## Linear regression                                      Number of obs =   2,021
##                                                        Replications  =   1,000
##                                                        Wald chi2(3)  =  122.77
##                                                        Prob > chi2   =  0.0000
##                                                        R-squared     =  0.0652
##                                                        Adj R-squared =  0.0638
##                                                        Root MSE      = 13.0608
##
## ------------------------------------------------------------------------------
```

```
##              |   Observed   Bootstrap                         Normal-based
##          HDL | coefficient  std. err.      z    P>|z|     [95% conf. interval]
## -------------+------------------------------------------------------------
##          BMI |  -.4036859   .0548886    -7.35   0.000    -.5112656   -.2961062
##          age |   .2086808   .0430616     4.85   0.000     .1242817      .29308
##     drinkany |   4.502504   .5957828     7.56   0.000     3.334792    5.670217
##        _cons |   46.68225   3.455292    13.51   0.000     39.91001     53.4545
## ------------------------------------------------------------------------------
##
## . estat bootstrap, normal
##
## Linear regression                               Number of obs     =      2,021
##                                                 Replications      =       1000
##
## ------------------------------------------------------------------------------
##              |    Observed               Bootstrap
##          HDL | coefficient       Bias    std. err.  [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.40368587  -.0038854    .0548886   -.5112656  -.2961062   (N)
##          age |   .20868084  -.0009771   .04306157    .1242817     .29308   (N)
##     drinkany |   4.5025044   .0024019   .59578279    3.334792   5.670217   (N)
##        _cons |   46.682253   .1759177   3.4552918    39.91001    53.4545   (N)
## ------------------------------------------------------------------------------
## Key: N: Normal
##
## . ** only percentile (R=1000)
## . bootstrap, reps(1000)  seed(12345): regress HDL BMI age drinkany
## (running regress on estimation sample)
##
## Bootstrap replications (1,000)
## ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
## ..................................................    50
## ..................................................   100
## ..................................................   150
## ..................................................   200
## ..................................................   250
## ..................................................   300
## ..................................................   350
## ..................................................   400
## ..................................................   450
## ..................................................   500
```

```
## ..................................................   550
## ..................................................   600
## ..................................................   650
## ..................................................   700
## ..................................................   750
## ..................................................   800
## ..................................................   850
## ..................................................   900
## ..................................................   950
## .................................................. 1,000
##
## Linear regression                                      Number of obs =   2,021
##                                                        Replications  =   1,000
##                                                        Wald chi2(3)  =  122.77
##                                                        Prob > chi2   =  0.0000
##                                                        R-squared     =  0.0652
##                                                        Adj R-squared =  0.0638
##                                                        Root MSE      = 13.0608
##
## ------------------------------------------------------------------------------
##              |   Observed   Bootstrap                         Normal-based
##          HDL | coefficient  std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.4036859   .0548886    -7.35   0.000    -.5112656   -.2961062
##          age |   .2086808   .0430616     4.85   0.000     .1242817      .29308
##     drinkany |   4.502504   .5957828     7.56   0.000     3.334792    5.670217
##        _cons |   46.68225   3.455292    13.51   0.000     39.91001     53.4545
## ------------------------------------------------------------------------------
##
## . estat bootstrap, percentile
##
## Linear regression                                Number of obs     =      2,021
##                                                  Replications      =       1000
##
## ------------------------------------------------------------------------------
##              |    Observed               Bootstrap
##          HDL | coefficient       Bias    std. err.  [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.40368587  -.0038854    .0548886   -.5188768  -.2992493   (P)
##          age |   .20868084  -.0009771   .04306157     .117856   .2910675   (P)
##     drinkany |   4.5025044   .0024019   .59578279    3.368253   5.692966   (P)
```

```
##        _cons |   46.682253   .1759177   3.4552918    40.42489   54.07221   (P)
## -------------------------------------------------------------------------------
## Key: P: Percentile
##
## . ** only BCa (R=1000), slightly DIFFERENT command
## . bootstrap, bca reps(1000)  seed(12345): regress HDL BMI age drinkany
## (running regress on estimation sample)
##
## Jackknife replications (2,021)
## ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
## .................................................    50
## .................................................   100
## .................................................   150
## .................................................   200
## .................................................   250
## .................................................   300
## .................................................   350
## .................................................   400
## .................................................   450
## .................................................   500
## .................................................   550
## .................................................   600
## .................................................   650
## .................................................   700
## .................................................   750
## .................................................   800
## .................................................   850
## .................................................   900
## .................................................   950
## ................................................. 1,000
## ................................................. 1,050
## ................................................. 1,100
## ................................................. 1,150
## ................................................. 1,200
## ................................................. 1,250
## ................................................. 1,300
## ................................................. 1,350
## ................................................. 1,400
## ................................................. 1,450
## ................................................. 1,500
## ................................................. 1,550
```

```
## .................................................. 1,600
## .................................................. 1,650
## .................................................. 1,700
## .................................................. 1,750
## .................................................. 1,800
## .................................................. 1,850
## .................................................. 1,900
## .................................................. 1,950
## .................................................. 2,000
## ..................
##
## Bootstrap replications (1,000)
## ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
## ..................................................    50
## ..................................................   100
## ..................................................   150
## ..................................................   200
## ..................................................   250
## ..................................................   300
## ..................................................   350
## ..................................................   400
## ..................................................   450
## ..................................................   500
## ..................................................   550
## ..................................................   600
## ..................................................   650
## ..................................................   700
## ..................................................   750
## ..................................................   800
## ..................................................   850
## ..................................................   900
## ..................................................   950
## .................................................. 1,000
##
## Linear regression                                 Number of obs =   2,021
##                                                   Replications  =   1,000
##                                                   Wald chi2(3)  =  122.77
##                                                   Prob > chi2   =  0.0000
##                                                   R-squared     =  0.0652
##                                                   Adj R-squared =  0.0638
##                                                   Root MSE      = 13.0608
##
```

```
## ------------------------------------------------------------------------------
##              |   Observed   Bootstrap                         Normal-based
##          HDL | coefficient  std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.4036859   .0548886    -7.35   0.000    -.5112656   -.2961062
##          age |   .2086808   .0430616     4.85   0.000     .1242817      .29308
##     drinkany |   4.502504   .5957828     7.56   0.000     3.334792    5.670217
##        _cons |   46.68225   3.455292    13.51   0.000     39.91001     53.4545
## ------------------------------------------------------------------------------
##
## . estat bootstrap
##
## Linear regression                               Number of obs     =      2,021
##                                                 Replications      =       1000
##
## ------------------------------------------------------------------------------
##              |    Observed               Bootstrap
##          HDL | coefficient       Bias    std. err.  [95% conf. interval]
## -------------+----------------------------------------------------------------
##          BMI |  -.40368587  -.0038854    .0548886    -.507547  -.2924751  (BC)
##          age |   .20868084  -.0009771   .04306157    .1173203   .2900742  (BC)
##     drinkany |   4.5025044   .0024019   .59578279    3.405528   5.748656  (BC)
##        _cons |   46.682253   .1759177   3.4552918    39.84275   53.29212  (BC)
## ------------------------------------------------------------------------------
## Key: BC: Bias-corrected
```

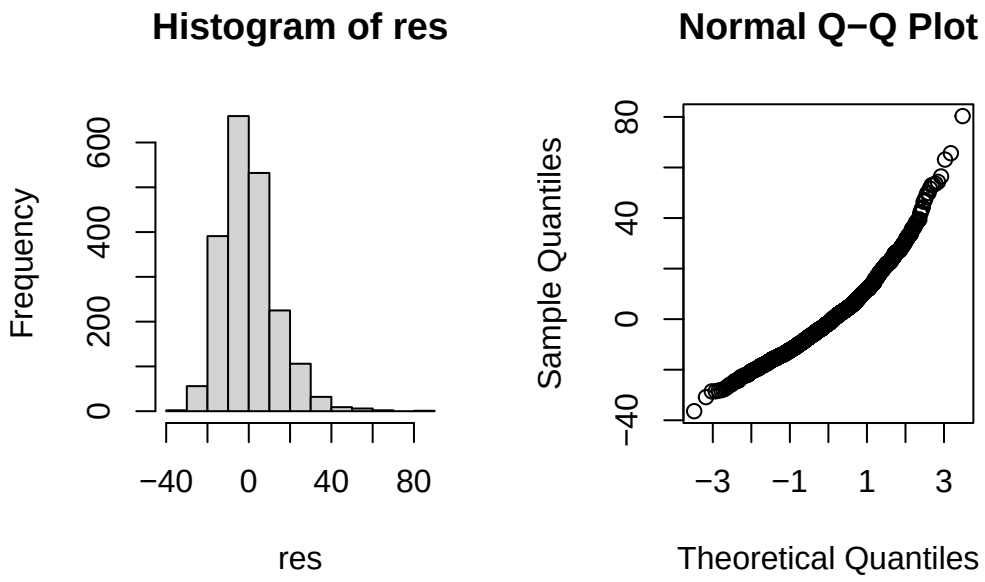
As you can see that all 3 bootstrap 95% CIs are similar to each other. Also, they are similar to the ones reported using the LS approach as reported in 1); we can therefore be confident that we don't have any particular issue with the standard analysis.

R code and output

1) standard regression of *HDL* on *BMI*, *age* and *drinkany* after removing diabetic patients. Examination of the residuals.

```
require(haven)
hers<-read_dta("hersdata.dta")
hers<-data.frame(hers)

# keep only the relevant variables and delete missing data
hers.nondiab<-hers[hers$diabetes==0,]
hers1<-cbind(hers.nondiab$HDL,hers.nondiab$age,hers.nondiab$BMI,hers.nondiab$drinkany)
```

```r
colnames(hers1)<-c("HDL","age","BMI","drinkany")
hers1<-data.frame(hers1)
hers2<-na.omit(hers1)
# 2032 --> 2021 observations

# standard analysis and residuals plots
out<-lm(HDL ~ BMI + age + drinkany, data=hers2)
res<-residuals(out)
par(mfrow=c(1,2))
hist(res)
qqnorm(res)
```




```r
# standard 95\% CI
confint(out)
##                  2.5 %     97.5 %
## (Intercept) 39.6773895 53.6871162
## BMI         -0.5156793 -0.2916924
## age          0.1228974  0.2944643
## drinkany     3.3492220  5.6557867
```

We clearly seem some upward curvature in the residuals. Given the large sample size, normality

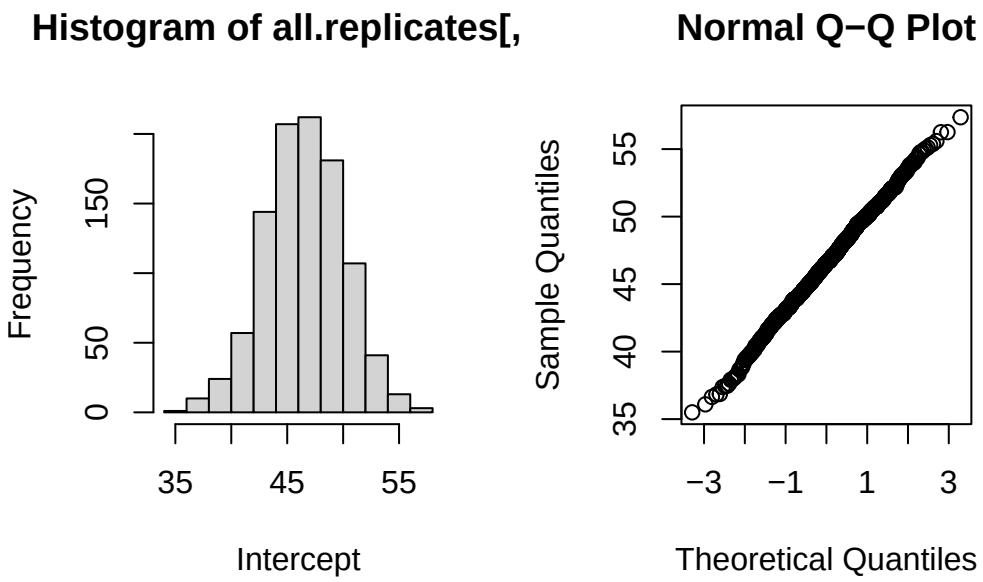
is not that critical but we are going to check that inference is indeed valid using the bootstrap.

2) Read Part A) of the code, run it and draw a histogram of the bootstrap samples (or replicates) for each of the coefficients.

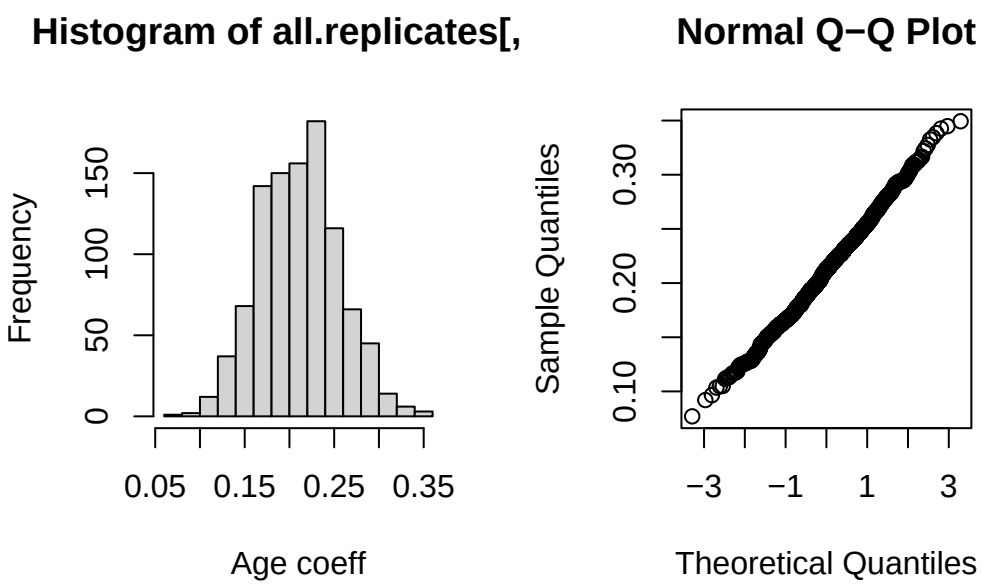
```
set.seed(1001)
R=1000
n=dim(hers2)[1]
all.replicates<-NULL
for(r in 1:R){
  # generate bootstrap sample by resampling the data
  hers2.r=hers2[sample(nrow(hers2), n,replace = TRUE), ]
  # fitted model (based on the bootstrap sample)
  out.r<-lm(HDL~age+BMI+drinkany,data=hers2.r)
  # store all coefficients in all.replicates
  all.replicates=rbind(all.replicates,out.r$coeff)
}

# all.replicates is a matrix Rx4 (since we have R replicates
# and 4 coefficients in the model)
dim(all.replicates)
## [1] 1000    4
head(all.replicates)
##      (Intercept)       age        BMI drinkany
## [1,]    46.75842 0.1993413 -0.3715308 4.097484
## [2,]    47.22997 0.2388648 -0.5078162 4.807770
## [3,]    49.87557 0.1395762 -0.3583760 4.899443
## [4,]    45.82771 0.2111516 -0.3688366 4.064384
## [5,]    51.28788 0.1659410 -0.4730509 4.431159
## [6,]    42.80404 0.2628449 -0.4131064 5.184175
```

The dataset all.replicates contains R=1000 replicates or bootstrap samples (1000x4 since we have 4 coefficients). Of course, you can change the number of replicates (e.g. use *R=3000* in the code above if you want 3000 replicates). R should be chosen large enough (at least 1000) for 95% confidence intervals. Histograms and normality plots follow easily.

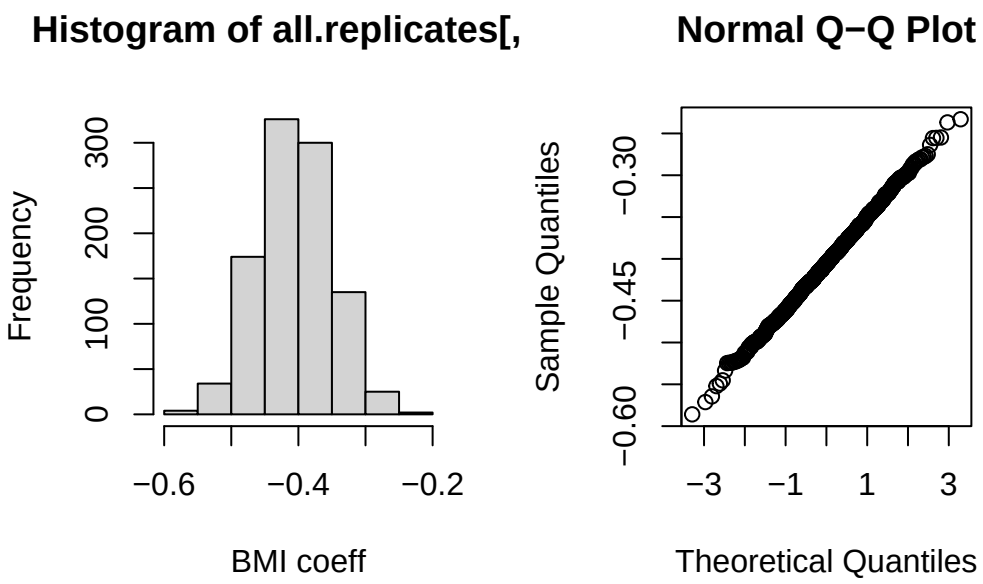
```
par(mfrow=c(1,2))
# intercept
hist(all.replicates[,1],xlab="Intercept")
qqnorm(all.replicates[,1])
```



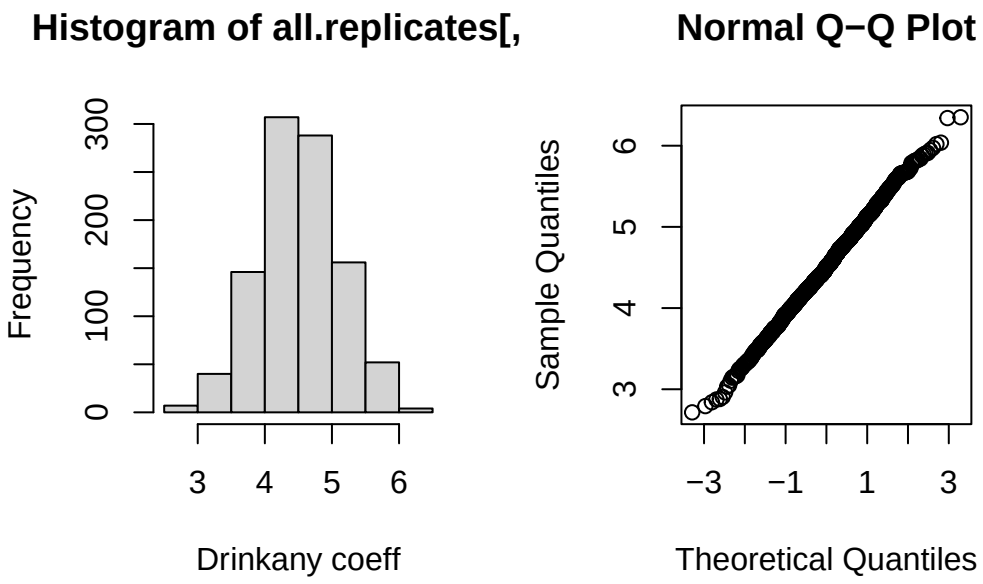

```
# age coefficient
hist(all.replicates[,2],xlab="Age coeff")
qqnorm(all.replicates[,2])
```




```
# BMI coefficient
hist(all.replicates[,3],xlab="BMI coeff")
qqnorm(all.replicates[,3])
```




```
# drinkany coefficien
hist(all.replicates[,4],xlab="Drinkany coeff")
qqnorm(all.replicates[,4])
```



3) direct calculation of the percentile 95% CI using a one-line command.

```
# intercept
quantile(all.replicates[,1], c(0.025,0.975))
##     2.5%    97.5%
## 39.56777 53.35830
# age
quantile(all.replicates[,2], c(0.025,0.975))
##      2.5%     97.5%
## 0.1268630 0.2978618
# BMI
quantile(all.replicates[,3], c(0.025,0.975))
##       2.5%      97.5%
## -0.5114304 -0.2988600
# drinkany
quantile(all.replicates[,4], c(0.025,0.975))
##     2.5%    97.5%
## 3.334052 5.673295
```

You simply calculate the 0.025th and 0.975th quantile for each coefficient (i.e. each column of all.replicates) to get the required percentile CIs.

4) Use the R library *boot* to avoid having to do the resampling "by hand''. Use Part B)

and compare to the standard analysis. You can decide to display all 3 types or only one. Note that we used here R=3000 replicates, the BCa appraoch had numerical issues with R=1000. If this happens to you increase the number of replicates.

```
library(boot)

# function collecting the coefficients; in general this function
# computes the statistic we want to bootstrap.
coeff<- function(data, indices){
  data <- data[indices,] # select obs. in bootstrap sample
  mod <- lm(HDL~age+BMI+drinkany, data=data) # modify formula here
  coefficients(mod) # return coefficient vector
}

# NB: R doc says or parametric bootstrap (i.e. the one we are using)
#  "the first argument to statistic must be the data"
#  " The second will be a vector of indices, frequencies or weights
#   which define the bootstrap sample".

# LS-based 95% CI to compare to

out<-lm(HDL~age+BMI+drinkany,data=hers2)
confint(out)
##                  2.5 %     97.5 %
## (Intercept) 39.6773895 53.6871162
## age          0.1228974  0.2944643
## BMI         -0.5156793 -0.2916924
## drinkany     3.3492220  5.6557867

set.seed(1001)
B = boot(data=hers2,statistic=coeff,R=3000)
# various 95% CI for the BMI coefficient (index=3)
# you can also get all the other ones by changing the index (e.g. index=2 for age)

# normal
boot.ci(B,index=3,type="norm")
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = B, type = "norm", index = 3)
##
```

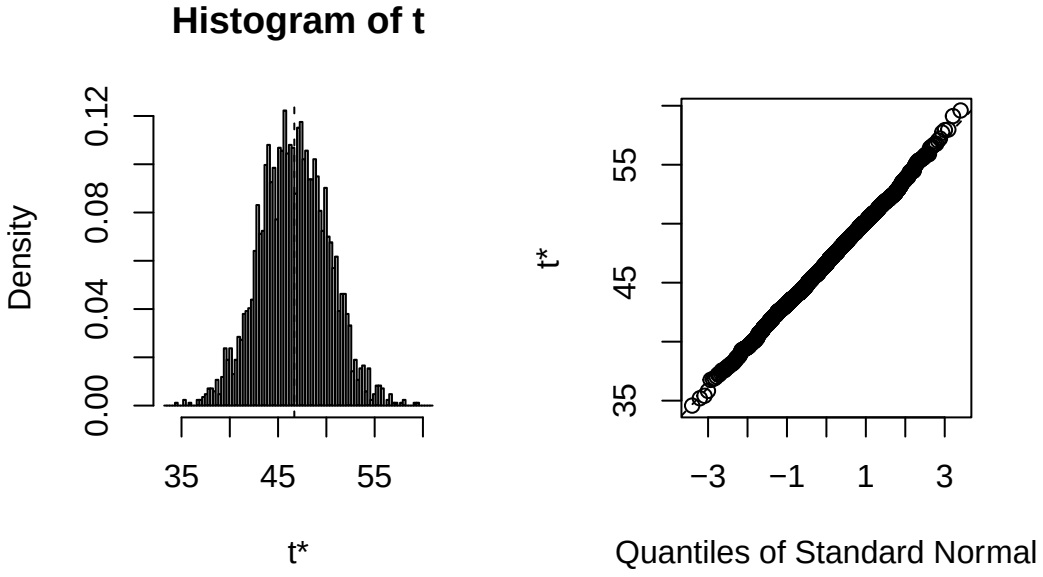
```
## Intervals :
## Level      Normal
## 95%   (-0.5113, -0.2958 )
## Calculations and Intervals on Original Scale
# percentile
boot.ci(B,index=3,type="perc")
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = B, type = "perc", index = 3)
##
## Intervals :
## Level     Percentile
## 95%   (-0.5179, -0.3000 )
## Calculations and Intervals on Original Scale
# BCa
boot.ci(B,index=3,type="bca")
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = B, type = "bca", index = 3)
##
## Intervals :
## Level       BCa
## 95%   (-0.5181, -0.3001 )
## Calculations and Intervals on Original Scale
# to get all 3 types in one command
boot.ci(B,index=3,type=c("norm","perc", "bca"))
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = B, type = c("norm", "perc", "bca"), index = 3)
##
## Intervals :
## Level      Normal              Percentile            BCa
## 95%   (-0.5113, -0.2958 )   (-0.5179, -0.3000 )   (-0.5181, -0.3001 )
## Calculations and Intervals on Original Scale
```

The results are very similar to the ones obtained by direct calculation (tiny differences), e.g. the

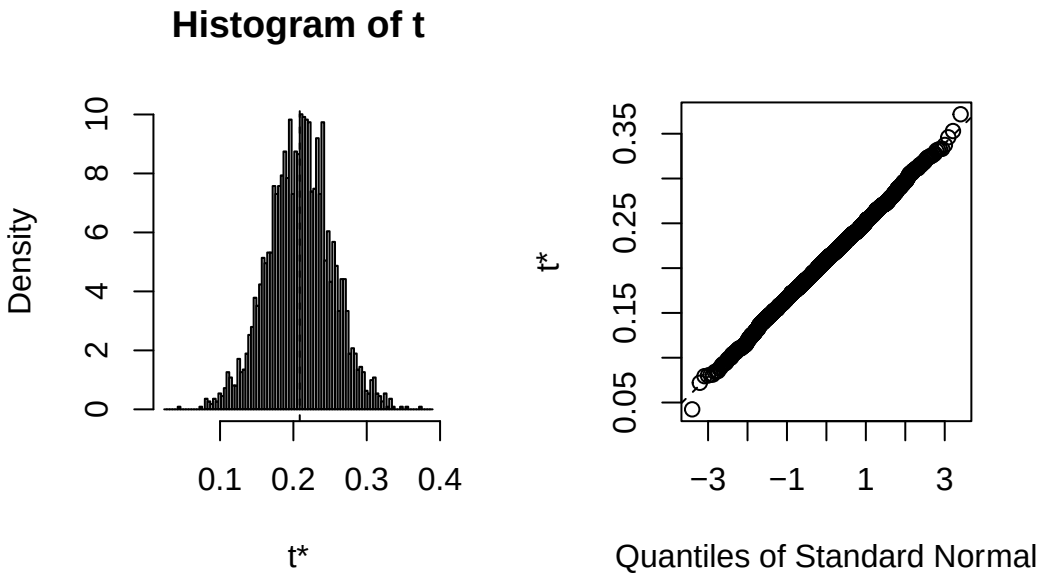
95% CI for the BMI coefficient is (-.519 ; -.300) using the percentile and Bca approaches. All three bootstrap 95% CIs are similar to each other for all coefficients. Also, we don't see substantial differences with the standard 95% cIs (i.e. the ones reported using the LS approach). We can be confident that we don't have any particular issue with the standard analysis.

Some plots (histogram, normal probabilit plots) of the bootstrap samples can be easily produced using the output B produced by *boot()*:

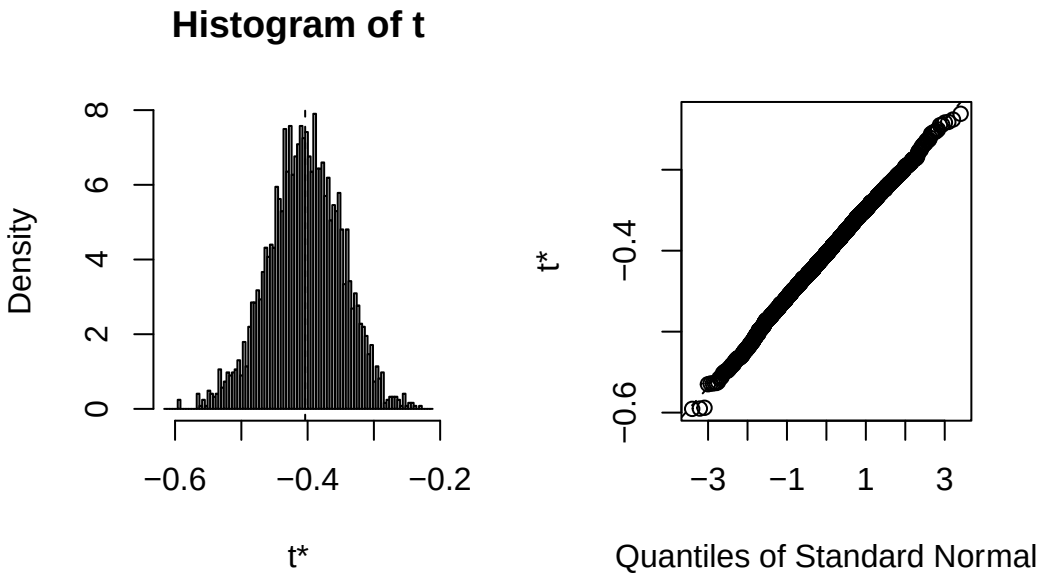
```
# intercept
plot(B, index=1)
```



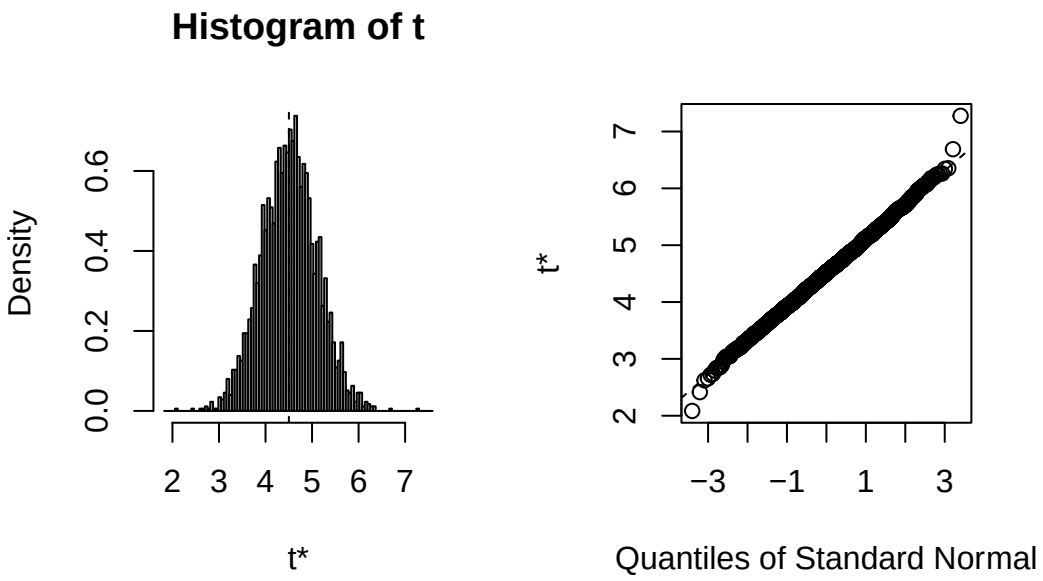

```
# x1=age
plot(B, index=2)
```




```
# x2=BMI
plot(B, index=3)
```

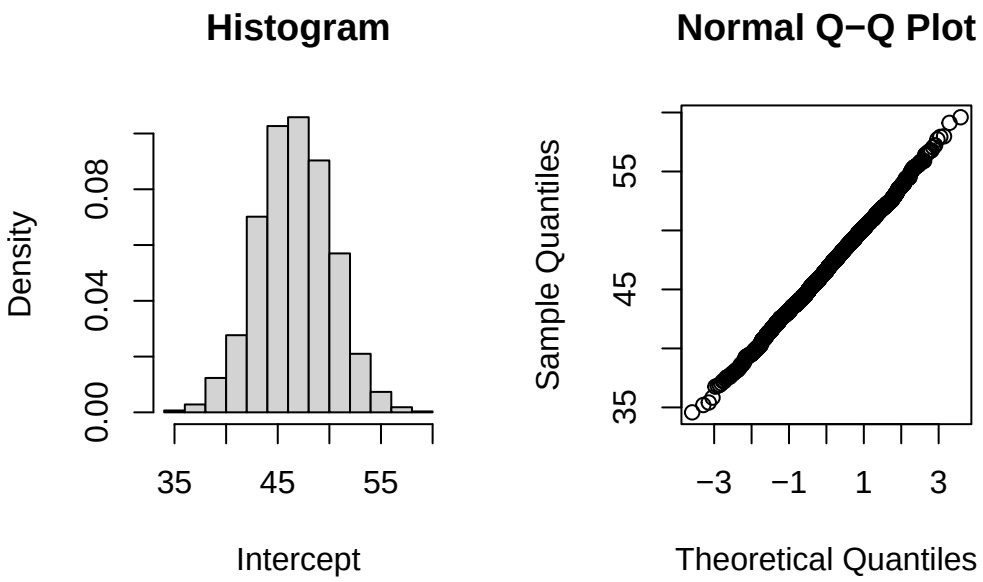


```
# x3=drinkany
plot(B, index=4)
```

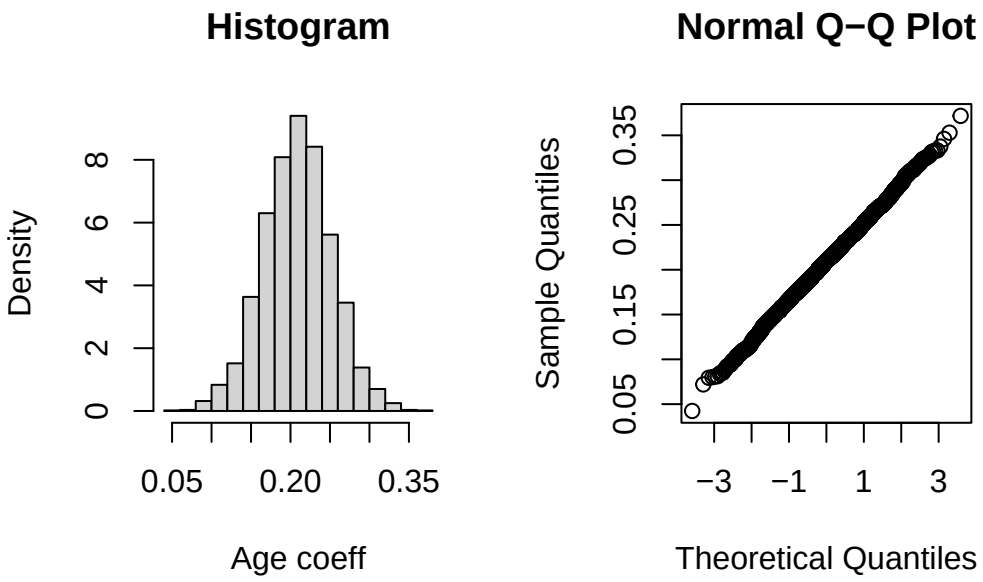


Nicer plots (you can control the labels, titles etc) can be produced using this code:

```
all<-B$t
par(mfrow=c(1,2))
hist(all[,1],main="Histogram",xlab="Intercept",prob=TRUE)
qqnorm(all[,1])
```




```
hist(all[,2],main="Histogram",xlab="Age coeff",prob=TRUE)
qqnorm(all[,2])
```



```
# etc
# hist(all[,3],main="Histogram",xlab="BMI coeff",prob=TRUE)
# qqnorm(all[,3])
# hist(all[,4],main="Histogram",xlab="drinkany coeff",prob=TRUE)
# qqnorm(all[,4])
```

Note that a somehow slower version can be obtained using a code like this (again illustrated on BMI using the percentile method). It is slower but possibly more "natural' ' since we bootstrap a statistic and provide a dataset and a formula for its calcualtion.

```
coeff<- function(data, indices, formula){
  data <- data[indices,]
  mod <- lm(formula=formula, data=data)
  coefficients(mod)
}

set.seed(1001)
B = boot(data=hers2,statistic=coeff,R=3000, formula=HDL~age+BMI+drinkany)
boot.ci(B,index=3,type="perc")
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = B, type = "perc", index = 3)
##
## Intervals :
## Level     Percentile
## 95%   (-0.5179, -0.3000 )
## Calculations and Intervals on Original Scale
```