Week 11 - Exercises-Solutions

Exercise solutions

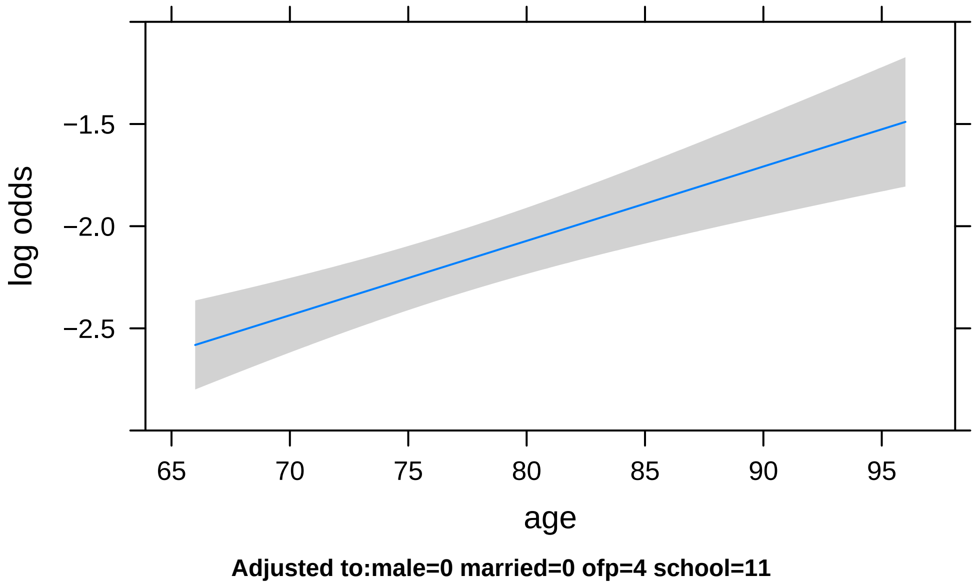
Week 11

Investigation

R code and output

1) initial model with covariates (model0) and AIC

```
library(rms)
## Loading required package: Hmisc
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
## The following objects are masked from 'package:base':
##
##     format.pval, units
medcare<- read.csv("medcare.csv")
medcare<-data.frame(medcare)
medcare$age<-medcare$age*10
ddist <- datadist(medcare)
options(datadist='ddist')
model0 <- lrm(healthpoor  ~  age + male + married + ofp + school, data=medcare)
plot(Predict(model0, age))
```



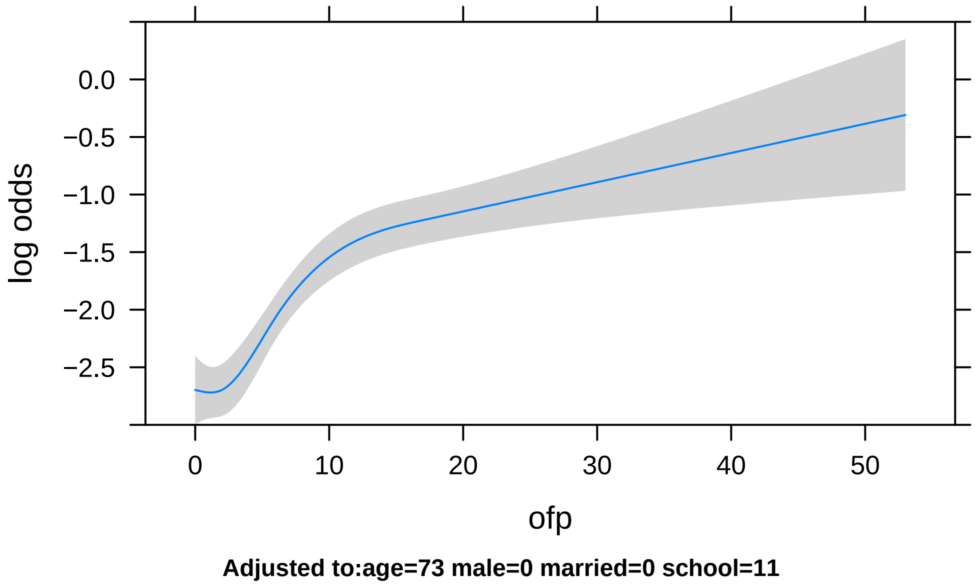

```
AIC(model0)
## [1] 3095.937
```

Only *age*, *ofp* and *school* are significant in this model that is the standard model without splines which acts as a starting point. AIC=3095.9 for this model. If you use the *plot(Predict(model0, age))* command after this fit you get a straight line since we did not include a spline in age.

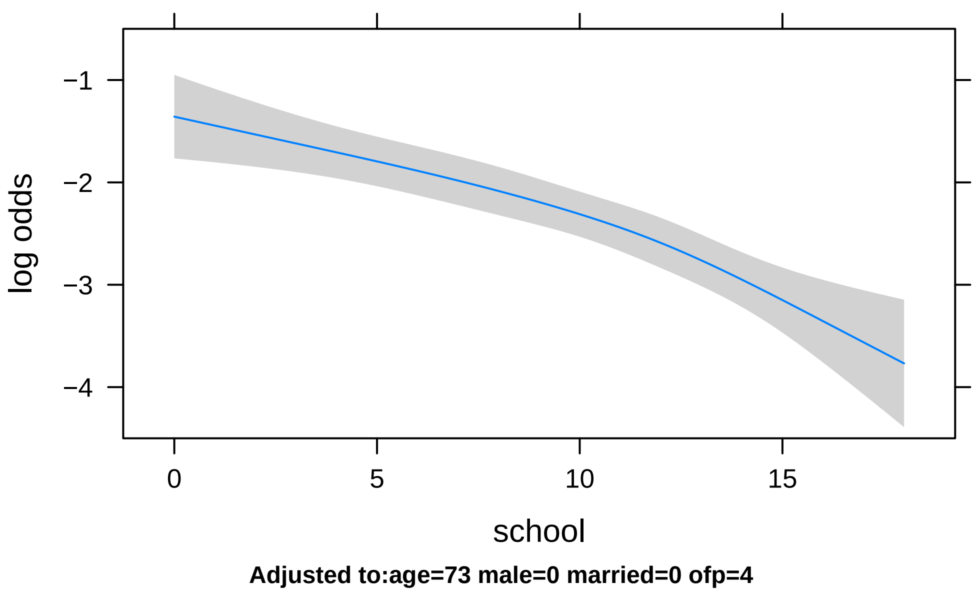
2) model with RCS(4) in *ofc* and *school*(model1) and AIC. Are splines necessary?

```
dist <- datadist(medcare)
options(datadist='ddist')
model1 <- lrm(healthpoor  ~  age + male + married + rcs(ofp,4) + rcs(school,4), data=medca
model1
## Logistic Regression Model
##
## lrm(formula = healthpoor ~ age + male + married + rcs(ofp, 4) +
##     rcs(school, 4), data = medcare)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes           Indexes
## Obs          4406    LR chi2      288.35    R2          0.119    C        0.711
##  0           3852    d.f.              9    R2(9,4406)0.061    Dxy      0.423
```

```
## 1             554    Pr(> chi2) <0.0001    R2(9,1453)0.175    gamma    0.423
## max |deriv| 4e-08                           Brier     0.102    tau-a    0.093
##
##           Coef    S.E.   Wald Z Pr(>|Z|)
## Intercept -4.2795 0.6204 -6.90  <0.0001
## age        0.0365 0.0075  4.90  <0.0001
## male       0.0168 0.1080  0.16  0.8765
## married   -0.0128 0.1088 -0.12  0.9066
## ofp       -0.0295 0.0793 -0.37  0.7096
## ofp'       2.1258 0.8687  2.45  0.0144
## ofp''     -3.5258 1.3579 -2.60  0.0094
## school    -0.0871 0.0323 -2.70  0.0069
## school'   -0.0400 0.0780 -0.51  0.6079
## school''  -0.0975 0.4821 -0.20  0.8397
AIC(model1)
## [1] 3064.379
anova(model1)
##                 Wald Statistics          Response: healthpoor
##
##  Factor          Chi-Square d.f. P
##  age              24.00     1    <.0001
##  male              0.02     1    0.8765
##  married           0.01     1    0.9066
##  ofp             161.72     3    <.0001
##   Nonlinear       35.22     2    <.0001
##  school           91.59     3    <.0001
##   Nonlinear        3.61     2    0.1641
##  TOTAL NONLINEAR  39.37     4    <.0001
##  TOTAL           263.85     9    <.0001
plot(Predict(model1,ofp))
```




```
plot(Predict(model1,school))
```



A spline in *ofp* is clearly needed (p<0.0001), with the log-odds of being in poor heath increasing markedly from 2 to 10 and less steeply after that. Note that a 1-2 visits to the doctor's don't seem to increase the odds of a poor outcome. A slight downward curvature is observed in the association with *school*, years of education, but there is no evidence that the spline is *school* is needed (p=0.16). Note that the plots have been drawn for other covariates set at their median values (by default) The AIC has been decreased subtantially compared with model0's, AIC=3064.4. We definetely need to keep a spline in *ofp* in the model (we could play around we the number of knots, their location but this would be further refinement). It's not so clear what do do with *school* since there is this apparent curvature. Options are: 1) go back to a simpler model with a linear term in *school*; 2) refine the modelling further to try and capture this curvature.

3) model with RCS(4) in *ofc* and a quadratic term in *school* (model2) and AIC.

```
##medcare$school2<-medcare$school^2
dist <- datadist(medcare)
options(datadist='ddist')
model2 <- lrm(healthpoor  ~  age + male + married + rcs(ofp,4) + poly(school,2,raw=TRUE),
model2
## Logistic Regression Model
##
## lrm(formula = healthpoor ~ age + male + married + rcs(ofp, 4) +
##     poly(school, 2, raw = TRUE), data = medcare)
##
##                       Model Likelihood     Discrimination    Rank Discrim.
##                             Ratio Test            Indexes          Indexes
## Obs           4406    LR chi2      290.10     R2          0.120    C        0.712
##  0            3852    d.f.              8    R2(8,4406)0.062    Dxy      0.423
##  1             554    Pr(> chi2) <0.0001    R2(8,1453)0.176    gamma    0.424
## max |deriv| 4e-09                            Brier      0.102    tau-a    0.093
##
##           Coef    S.E.   Wald Z Pr(>|Z|)
## Intercept -4.3877 0.6258 -7.01  <0.0001
## age        0.0363 0.0074  4.90  <0.0001
## male       0.0175 0.1079  0.16  0.8713
## married   -0.0113 0.1088 -0.10  0.9169
## ofp       -0.0311 0.0793 -0.39  0.6951
## ofp'       2.1401 0.8684  2.46  0.0137
## ofp''     -3.5476 1.3575 -2.61  0.0090
## 1         -0.0266 0.0431 -0.62  0.5369
## 2         -0.0056 0.0025 -2.30  0.0213
AIC(model2)
```

```
## [1] 3060.625
anova(model2)
##                 Wald Statistics          Response: healthpoor
##
##  Factor     Chi-Square d.f. P
##  age         24.03     1    <.0001
##  male         0.03     1    0.8713
##  married      0.01     1    0.9169
##  ofp        161.66     3    <.0001
##   Nonlinear  35.13     2    <.0001
##  school      92.55     2    <.0001
##  TOTAL      264.80     8    <.0001
```

There is now evidence that the quadratic term is necessary (ANOVA returns p<0.0001) for the global effect of the two ofp terms. You can also get a similar result by defining the quadratic term by hand in the dataset, fitting the model and computing a LRT testing whether these two terms are necessary. It's simpler to use *poly()* and *anova()*. The AIC has decreased further for this model (model2) since AIC=3060.6

4) What is the best model fitted so far based on the AIC (or BIC)?

Model2 is the better model due its smaller AIC if we consider this statistic to rank models. The command: *BIC(model0, model1, model2)* gives the corresponding BIC values, i.e. 3134.3, 3128.3 and 3118.1 favouring more neatly model2 (BIC=3118.1 is neatly smaller than the two other BIC's). So there seems to be evidence of small quadratic term as indicated first in the plot.

```
BIC(model0, model1, model2)
##        df      BIC
## model0  6 3134.282
## model1 10 3128.287
## model2  9 3118.142
```

5) smaller AIC/BIC? further refinements

We could try to play with the knots but a simple way to possibly reduce further the AIC/BIC is to remove the non-significant variables e.g. *married* and *male* yielding the following results:

```
dist <- datadist(medcare)
options(datadist='ddist')
model3 <- lrm(healthpoor  ~  age + rcs(ofp,4) + poly(school,2,raw=TRUE), data=medcare)
model3
```

```
## Logistic Regression Model
##
## lrm(formula = healthpoor ~ age + rcs(ofp, 4) + poly(school, 2,
##     raw = TRUE), data = medcare)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                              Ratio Test            Indexes          Indexes
## Obs            4406    LR chi2     290.07    R2         0.120    C        0.712
##  0             3852    d.f.             6    R2(6,4406)0.062    Dxy      0.423
##  1              554    Pr(> chi2) <0.0001    R2(6,1453)0.178    gamma    0.424
## max |deriv| 4e-09                            Brier      0.102    tau-a    0.093
##
##           Coef    S.E.   Wald Z Pr(>|Z|)
## Intercept -4.3926 0.6001 -7.32  <0.0001
## age        0.0365 0.0072  5.09  <0.0001
## ofp       -0.0316 0.0792 -0.40  0.6893
## ofp'       2.1438 0.8676  2.47  0.0135
## ofp''     -3.5528 1.3564 -2.62  0.0088
## 1         -0.0272 0.0429 -0.63  0.5271
## 2         -0.0056 0.0024 -2.30  0.0216
AIC(model3)
## [1] 3056.653
BIC(model3)
## [1] 3101.388
anova(model3)
##                Wald Statistics          Response: healthpoor
##
##  Factor     Chi-Square d.f. P
##  age         25.87     1    <.0001
##  ofp        162.27     3    <.0001
##   Nonlinear  35.16     2    <.0001
##  school      94.00     2    <.0001
##  TOTAL      264.79     6    <.0001
```

AIC=3056.7 and BIC=3101.4 have been further reduced suggesting that this more parsimonious model is preferable (unless there is external evidence to keep *married* and *male*, for instance due to their confounding effect in other studies). Such evidence is lacking so we may be happy to stick with model3 from a purely statistical perspective. We have not formally validated the model but using splines or polynomials is no substitute for validation. Often, we deal with outliers and influential observations prior to this sort of modelling.

6) Conclusions

There is no unique way to describe the different steps but here is one that starts by describing what we are trying to do, the different steps, what we found and describe the final model. We investigated the association between poor health (*poorhealth*) and various predictors, i.e. age , male, the number of physician office visits (*ofp*), years of education (*school*) using logistic regression. Since associations with continous covariates were not necessary linear (on the log-odds scale), we used restricted cubic splines and polynomials to relax this assumption. There was no enough evidence to suggest that the association with age was not linear but a spline in *ofp* was necessary. The log-odds of being in poor heath increases markedly with *ofp* from 2 to 10 and less steeply after that. The relationship of *poorhealth* with *school* (on the log-odds scale) is better captured by a quadratic polynomial displaying a faster decay with larger values of years of educations. Plots can be referred to to support that claim. The AIC/BIC confirmed that such a model was indeed preferable. A more parsimonious model (i.e. without the non-significant predictors *married* and *age*) is supported by a smaller AIC/BIC. You can also gives some ORs and 95% CIs for the linear association(s, only age if you keep the latter model. The OR for age is OR=exp(0.0365)=1.037, 95%CI=(1.023 ; 1.053) i.e. on average the odds increases by about 4%, 95% CI=(2.3% ; 5.2%) per additional year of age.

Stata code and output

1) initial model with covariates (model0) and AIC

```
use medcare.dta
replace age=age*10
logistic healthpoor age married male ofp school, coef
estat ic
## . use medcare.. replace age=age*10
## (4,406 real changes made)
##
## . logistic healthpoor age married male ofp school, coef
##
## Logistic regression                                     Number of obs =  4,406
##                                                         LR chi2(5)    = 248.79
##                                                         Prob > chi2   = 0.0000
## Log likelihood = -1541.9687                             Pseudo R2     = 0.0746
##
## ------------------------------------------------------------------------------
##   healthpoor | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          age |   .0363948   .0072894     4.99   0.000     .0221079    .0506817
##      married |   .0056677   .1080392     0.05   0.958    -.2060853    .2174207
##         male |  -.0418784   .1066731    -0.39   0.695    -.2509538     .167197
##          ofp |    .064116   .0057524    11.15   0.000     .0528415    .0753905
##       school |  -.1209181   .0123262    -9.81   0.000     -.145077   -.0967592
```

```
##        _cons |  -3.909815   .5904464    -6.62   0.000    -5.067069   -2.752562
## ------------------------------------------------------------------------------
##
## . estat ic
##
## Akaike's information criterion and Bayesian information criterion
##
## -----------------------------------------------------------------------------
##        Model |          N   ll(null)  ll(model)      df         AIC        BIC
## -------------+---------------------------------------------------------------
##            . |      4,406  -1666.363  -1541.969       6    3095.937   3134.282
## -----------------------------------------------------------------------------
## Note: BIC uses N = number of observations. See [R] BIC note.
```

Only *age*, *ofp* and *school* are significant in this model that is the standard model without splines which acts as a starting point. AIC=3095.9 for this model.

2) model with RCS(4) in *ofc* and *school*(model1) and AIC. Are splines necessary?

```
clear
use medcare.dta
replace age=age*10
mkspline2 ofpspl = ofp, cubic nknots(4)
mkspline2 schoolspl = school, cubic nknots(4)
logistic healthpoor age married male ofpspl* schoolspl*, coef
** splines for school) (on the logit scale)
adjustrcspline, at(age=73 married=1 male=0 ofp=4) custominvlink("xb()") ytitle("log-odds")
** NB: caution with the scale - default= proba
** logit scale via the option custominvlink("xb()"
estat ic
test ofpspl2 ofpspl3
test schoolspl2 schoolspl3
** ----------------------------------------
** to get the second plot refit the model
** ----------------------------------------
clear
use medcare.dta
replace age=age*10
mkspline2 schoolspl = school, cubic nknots(4)
mkspline2 ofpspl = ofp, cubic nknots(4)
quiet logistic healthpoor age married male ofpspl* schoolspl*, coef
** splines for ofp (on the logit scale)
```

```
adjustrcspline if ofp <=50, at(age=73 married=1 male=0 school=11) custominvlink("xb()") yt
**logit scale via the option custominvlink("xb()")
**
** figures will be displayed when you run the code.
## . cl. use medcare.dta
##
## . replace age=age*10
## (4,406 real changes made)
##
## . mkspline2 ofpspl = ofp, cubic nknots(4)
## command mkspline2 is unrecognized
## r(199);
##
## end of do-file
## r(199);
```

A spline in *ofp* is clearly needed (p<0.0001), with the log-odds of being in poor heath increasing markedly from 2 to 10 and less steeply after that. Note that a 1-2 visits to the doctor's don't seem to increase the odds of a poor outcome. A slight downward curvature is observed in the association with *school*, years of education, but there is no evidence that the spline is *school* is needed (p=0.16). Note that the plots have been drawn for other covariates set at their median values (by default) The AIC has been decreased subtantially compared with model0's, AIC=3064.4. We definetely need to keep a spline in *ofp* in the model (we could play around we the number of knots, their location but this would be further refinement). It's not so clear what do do with *school* since there is this apparent curvature. Options are: 1) go back to a simpler model with a linear term in *school*; 2) refine the modelling further to try and capture this curvature.

3) model with RCS(4) in *ofc* and a quadratic term in *school* (model2) and AIC.

```
clear
use medcare.dta
replace age=age*10
gen school2=school^2
mkspline2 ofpspl = ofp, cubic nknots(4)
mkspline2 schoolspl = school, cubic nknots(4)
logistic healthpoor age married male ofpspl* school school2, coef
estat ic
test ofpspl2 ofpspl3
test school school2
## . cl. use medcare.dta
##
```

```
## . replace age=age*10
## (4,406 real changes made)
##
## . gen school2=school^2
##
## . mkspline2 ofpspl = ofp, cubic nknots(4)
## command mkspline2 is unrecognized
## r(199);
##
## end of do-file
## r(199);
```

There is now evidence that the quadratic term is necessary (test 2df returns p<0.0001) for the global effect of the two ofp terms. The AIC has decreased further for this model (model2) since AIC=3060.6

4) What is the best model fitted so far based on the AIC (or BIC)?

Model2 is the better model due its smaller AIC if we consider this statistic to rank models. The command: *estat ic* after each model fit gives the corresponding BIC values, i.e. 3134.3, 3128.3 and 3118.1 favouring more neatly model2 (BIC=3118.1 is neatly smaller than the two other BIC's). So there seems to be evidence of small quadratic term as indicated first in the plot.

5) smaller AIC/BIC? further refinements

We could try to play with the knots but a simple way to possibly reduce further the AIC/BIC is to remove the non-significant variables e.g. *married* and *male* yielding the following results:

```
clear
use medcare.dta
replace age=age*10
gen school2=school^2
mkspline2 ofpspl = ofp, cubic nknots(4)
mkspline2 schoolspl = school, cubic nknots(4)
logistic healthpoor age  ofpspl* school school2, coef
estat ic
test ofpspl2 ofpspl3
test school school2
** OR for age
lincom age, or
## . cl. use medcare.dta
##
```

```
## . replace age=age*10
## (4,406 real changes made)
##
## . gen school2=school^2
##
## . mkspline2 ofpspl = ofp, cubic nknots(4)
## command mkspline2 is unrecognized
## r(199);
##
## end of do-file
## r(199);
```

AIC=3056.7 and BIC=3101.4 have been further reduced suggesting that this more parsimonious model is preferable (unless there is external evidence to keep *married* and *male*, for instance due to their confounding effect in other studies). Such evidence is lacking so we may be happy to stick with model3 from a purely statistical perspective. We have not formally validated the model but using splines or polynomials is no substitute for validation. Often, we deal with outliers and influential observations prior to this sort of modelling.

6) Conclusions

There is no unique way to describe the different steps but here is one that starts by describing what we are trying to do, the different steps, what we found and describe the final model. We investigated the association between poor health (*poorhealth*) and various predictors, i.e. age , male, the number of physician office visits (*ofp*), years of education (*school*) using logistic regression. Since associations with continous covariates were not necessary linear (on the log-odds scale), we used restricted cubic splines and polynomials to relax this assumption. There was no enough evidence to suggest that the association with age was not linear but a spline in *ofp* was necessary. The log-odds of being in poor heath increases markedly with *ofp* from 2 to 10 and less steeply after that. The relationship of *poorhealth* with *school* (on the log-odds scale) is better captured by a quadratic polynomial displaying a faster decay with larger values of years of educations. Plots can be referred to to support that claim. The AIC/BIC confirmed that such a model was indeed preferable. A more parsimonious model (i.e. without the non-significant predictors *married* and *age*) is supported by a smaller AIC/BIC. You can also gives some ORs and 95% CIs for the linear association(s, only age if you keep the latter model. The OR for age is OR=exp(0.0365)=1.037, 95%CI=(1.02 ; 1.05) i.e. on average the odds increases by 3.7%, 95% CI=(2.2% ; 5.1%) per additional year of age.