Notes for laboratory session 4

1. Model Selection

Model selection in the linear regression model

We can decide if a variable should be excluded from a linear model by calculating the appropriate F-statistic. The calculation requires the residual sum of squares of the full model and the sub-model. Thus, we fit the full model…

```
. xi: reg  retplasm age i.sex i.smokstat quetelet i.vituse calories fat fiber
alcohol chol
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.smokstat        _Ismokstat_1-3      (naturally coded; _Ismokstat_1 omitted)
i.vituse          _Ivituse_1-3        (naturally coded; _Ivituse_3 omitted)

      Source |       SS       df       MS              Number of obs =     314
-------------+------------------------------           F( 12,   301) =    4.06
       Model |  1896984.44    12  158082.037           Prob > F      =  0.0000
    Residual |  11723197.4   301  38947.4997           R-squared     =  0.1393
-------------+------------------------------           Adj R-squared =  0.1050
       Total |  13620181.9   313   43514.958           Root MSE      =  197.35
------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   2.653472   .8756372     3.03   0.003     .9303267    4.376618
      _Isex_1 |    76.8363   37.37679     2.06   0.041     3.283403    150.3892
_Ismokstat_2 |   44.90691   25.13723     1.79   0.075    -4.560058    94.37388
_Ismokstat_3 |  -.6574155   36.25566    -0.02   0.986    -72.00408    70.68925
    quetelet |   1.581298   1.917623     0.82   0.410    -2.192347    5.354944
  _Ivituse_1 |   35.40501   27.26527     1.30   0.195    -18.24968    89.05969
  _Ivituse_2 |    27.8062   29.71094     0.94   0.350    -30.66125    86.27365
    calories |   .0758574   .0598645     1.27   0.206    -.0419486    .1936634
         fat |  -1.512089   .9335381    -1.62   0.106    -3.349177    .3249986
       fiber |  -4.207861   3.100573    -1.36   0.176    -10.30941    1.893684
     alcohol |   7.371856   2.602759     2.83   0.005     2.249949    12.49376
        chol |  -.0775529   .1048078    -0.74   0.460    -.2838016    .1286959
       _cons |   416.1679   83.85834     4.96   0.000      251.145    581.1907
------------------------------------------------------------------------------
```

SSE_{full}= …11723197.4

and then the sub-model (i.e. excluding the variable cholesterol) …

```
. xi: reg  retplasm age i.sex i.smokstat quetelet i.vituse calories fat fiber
alcohol
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.smokstat        _Ismokstat_1-3      (naturally coded; _Ismokstat_1 omitted)
i.vituse          _Ivituse_1-3        (naturally coded; _Ivituse_3 omitted)

      Source |       SS       df       MS              Number of obs =     314
-------------+------------------------------           F( 11,   302) =    4.38
       Model |  1875659.49    11  170514.499           Prob > F      =  0.0000
    Residual |  11744522.4   302  38889.1469           R-squared     =  0.1377
-------------+------------------------------           Adj R-squared =  0.1063
       Total |  13620181.9   313   43514.958           Root MSE      =  197.20

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   2.678755   .8743146     3.06   0.002     .9582353    4.399275
     _Isex_1 |   72.77019   36.94293     1.97   0.050     .0720315    145.4683
_Ismokstat_2 |    46.0355   25.07212     1.84   0.067    -3.302663    95.37367
_Ismokstat_3 |   .1518775     36.212     0.00   0.997    -71.10792    71.41168
    quetelet |   1.536417   1.915227     0.80   0.423    -2.232463    5.305297
  _Ivituse_1 |   36.63589   27.19409     1.35   0.179    -16.87799    90.14978
  _Ivituse_2 |   28.56312   29.67107     0.96   0.336    -29.82509    86.95134
    calories |   .0674277   .0587265     1.15   0.252    -.0481373    .1829927
         fat |  -1.592929   .9264287    -1.72   0.087    -3.416001    .2301444
       fiber |  -3.812586   3.051921    -1.25   0.213    -9.818308    2.193137
     alcohol |   7.534269   2.591544     2.91   0.004     2.434499    12.63404
       _cons |    412.199   83.62392     4.93   0.000     247.6397    576.7583
```

$SSE_{\text{sub-model}}$= …`11744522.4`

Then we can calculate the F-statistic using the formula:

$$\frac{SSE(X_p) - SSE(X_{p+1})}{SSE(X_{p+1}) \Big/ (n-p-1)} \sim F_{1,n-p-1}$$

Now we can calculate the upper-tail cumulative F distribution with 1 numerator and 314-12-1=301 denominator degrees of freedom using the STATA function `Ftail`.

```
. di Ftail(1,301,(11744522.4-11723197.4)/(11723197.4/301))
.45990464
```

Alternatively, we can use the STATA `test` command as follows:

```
. quietly xi: reg  retplasm age i.sex i.smokstat quetelet i.vituse calories fat
fiber alcohol chol

. test chol

 ( 1)  chol = 0.0

       F(  1,   301) =    0.55
            Prob > F =    0.4599
```

a) Which t-test is equivalent with the F-test above? Try to confirm this using STATA function `ttail` to calculate the relevant p-value.

Model selection in the GLM

In the GLM model the decision whether a variable should be excluded, can be based on a likelihood ratio test. The calculation requires the deviances (or the maximized log likelihoods) of the full and the sub-model respectively.

We fit the full model …

```
. xi: glm retplasm i.sex age i.smokstat i.vituse  quetelet calories fat fiber a
> lcohol  chol
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.smokstat        _Ismokstat_1-3      (naturally coded; _Ismokstat_1 omitted)
i.vituse          _Ivituse_1-3        (naturally coded; _Ivituse_3 omitted)

Iteration 0:   log likelihood = -2098.3936

Generalized linear models                         No. of obs      =       314
Optimization     : ML: Newton-Raphson             Residual df     =       301
                                                  Scale parameter =   38947.5
Deviance         =  11723197.42                   (1/df) Deviance =   38947.5
Pearson          =  11723197.42                   (1/df) Pearson  =   38947.5

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]
Standard errors  : OIM

Log likelihood   =  -2098.39358                   AIC             =  13.44837
BIC              =  11721466.85

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |    76.8363   37.37679     2.06   0.040     3.579146    150.0935
         age |   2.653472   .8756372     3.03   0.002     .9372552     4.36969
_Ismokstat_2 |   44.90691   25.13723     1.79   0.074     -4.36116    94.17499
_Ismokstat_3 |  -.6574155   36.25566    -0.02   0.986    -71.71721    70.40238
  _Ivituse_1 |   35.40501   27.26527     1.30   0.194    -18.03395    88.84396
  _Ivituse_2 |    27.8062   29.71094     0.94   0.349    -30.42617    86.03856
    quetelet |   1.581298   1.917623     0.82   0.410    -2.177174     5.33977
    calories |   .0758574   .0598645     1.27   0.205     -.041475    .1931897
         fat |  -1.512089   .9335381    -1.62   0.105     -3.34179     .317612
       fiber |  -4.207861   3.100573    -1.36   0.175    -10.28487    1.869151
     alcohol |   7.371856   2.602759     2.83   0.005     2.270543    12.47317
        chol |  -.0775529   .1048078    -0.74   0.459    -.2829723    .1278666
       _cons |   416.1679   83.85834     4.96   0.000     251.8086    580.5272
------------------------------------------------------------------------------
```

$\text{Deviance}_{\text{full}}$ = …

and then the sub-model …

```
. xi: glm retplasm i.sex age i.smokstat i.vituse  quetelet calories fat fiber a
> lcohol
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.smokstat        _Ismokstat_1-3      (naturally coded; _Ismokstat_1 omitted)
i.vituse          _Ivituse_1-3        (naturally coded; _Ivituse_3 omitted)

Iteration 0:   log likelihood = -2098.6789

Generalized linear models                         No. of obs      =       314
Optimization     : ML: Newton-Raphson             Residual df     =       302
                                                  Scale parameter =  38889.15
Deviance         =  11744522.37                   (1/df) Deviance =  38889.15
Pearson          =  11744522.37                   (1/df) Pearson  =  38889.15

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]
Standard errors  : OIM

Log likelihood   =  -2098.67891                   AIC             =  13.44381
BIC              =  11742786.06

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   72.77019   36.94293     1.97   0.049     .3633701     145.177
         age |   2.678755   .8743146     3.06   0.002     .9651303     4.39238
_Ismokstat_2 |    46.0355   25.07212     1.84   0.066     -3.10494    95.17595
_Ismokstat_3 |   .1518775     36.212     0.00   0.997    -70.82235     71.1261
  _Ivituse_1 |   36.63589   27.19409     1.35   0.178    -16.66354    89.93532
  _Ivituse_2 |   28.56312   29.67107     0.96   0.336     -29.5911    86.71735
    quetelet |   1.536417   1.915227     0.80   0.422     -2.21736    5.290193
    calories |   .0674277   .0587265     1.15   0.251    -.0476742    .1825296
         fat |  -1.592929   .9264287    -1.72   0.086    -3.408695    .2228384
       fiber |  -3.812586   3.051921    -1.25   0.212     -9.79424    2.169069
     alcohol |   7.534269   2.591544     2.91   0.004     2.454936     12.6136
       _cons |    412.199   83.62392     4.93   0.000     248.2991    576.0989
------------------------------------------------------------------------------
```

$\text{Deviance}_{\text{sub-model}}$ = …

The calculation of the log-likelihood statistic and its relevant p-value can be based on deviances…

```
. di chi2tail(1,(11744522.3740-11723197.4192)/((11723197.4192)/(301)))
.45932838
```

Likelihood ratio test using the `lrtest` command

```
. qui xi: glm retplasm i.sex age i.smokstat i.vituse  quetelet calories fat fiber alcohol
chol

. est store A

. qui xi: glm retplasm i.sex age i.smokstat i.vituse  quetelet calories fat fiber alcohol

. est store B

. lrtest A B,stats
(log-likelihoods of null models cannot be compared)

likelihood-ratio test                                 LR chi2(1)  =      0.57
(Assumption: B nested in A)                           Prob > chi2 =    0.4500

-------------------------------------------------------------------------------
Model        |   nobs    ll(null)   ll(model)     df         AIC          BIC
-------------+-----------------------------------------------------------------
           B |    314           .   -2098.679     12     4221.358    4266.351
           A |    314           .   -2098.394     13     4222.787    4271.529
-------------------------------------------------------------------------------
```

The easiest way to assess the impact of the factor cholesterol in the model is with the test command, which generates the Wald test.

```
.  quietly xi: glm retplasm i.sex age i.smokstat i.vituse  quetelet calories fa
> t fiber alcohol  chol
. test chol

 ( 1)  [retplasm]chol = 0

           chi2(  1) =    0.55
         Prob > chi2 =    0.4593
```

2. Model checking

In order to check the assumptions of the previous model we need to produce the predicted values and the residuals.

First we fit the model using the `regress` STATA command

```
. quietly xi:reg retplasm i.sex fat alcohol age
```

Then we generate the predicted values

```
. predict yhat
(option xb assumed; fitted values)
(1 missing value generated)
```

then the raw residuals…

```
. predict r, resid
(1 missing value generated)
```

the standardized residuals

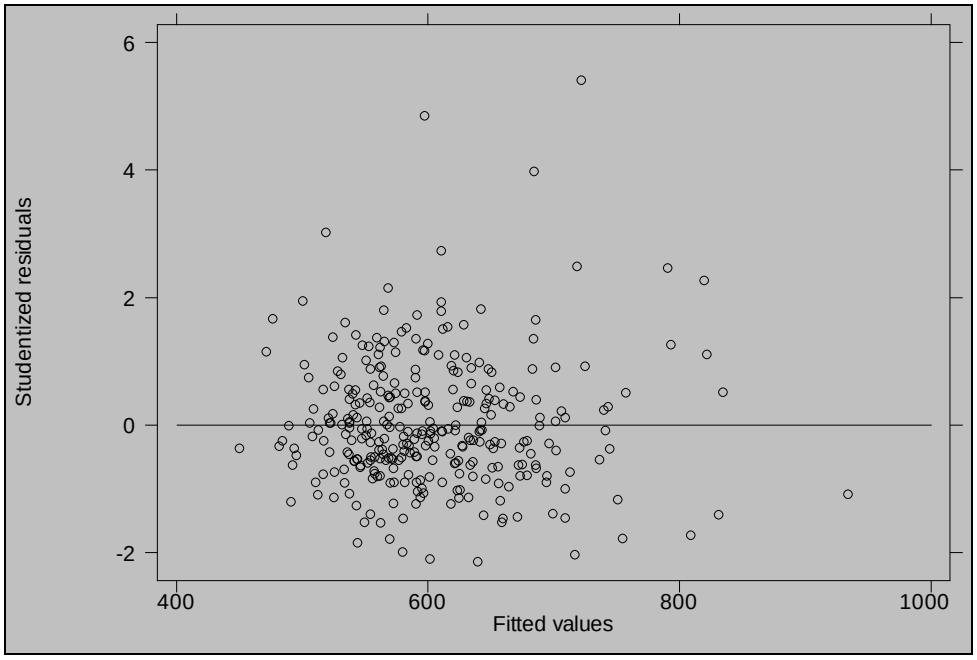
```
. predict rstan, rstand
(1 missing value generated)
```

and the Student's residuals

```
. predict rstud, rstud
(1 missing value generated)
```

We can now check the homoskedacity assumption by plotting the Studentized residuals versus the predicted values.

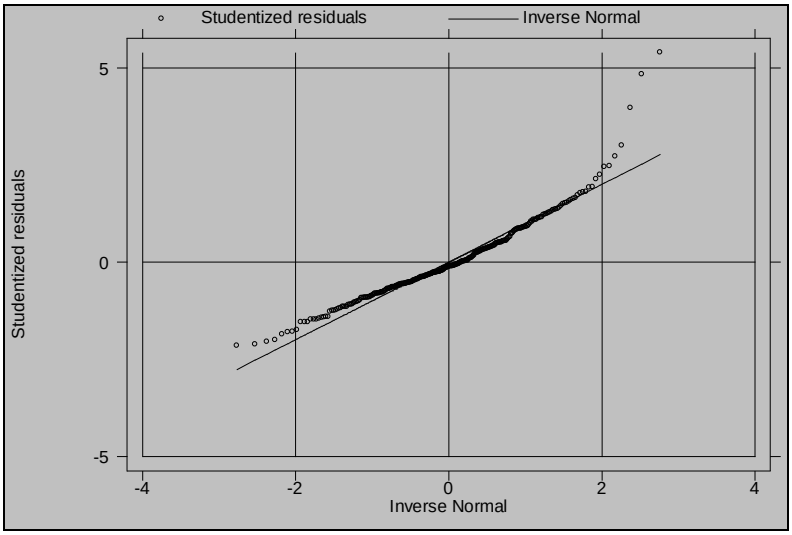
```
. sc rstud yhat,yline(0) xlab() ylab()
```



We see that there is no obvious problem with lack of homoskedasticity in these data. (Why?)

The assumption of Normality can be checked using the qnorm command in STATA as follows:

```
. qnorm rstud, xlab() ylab()
```



To formally test the hypothesis of normality, we can use the Shapiro-Wilks test as follows:

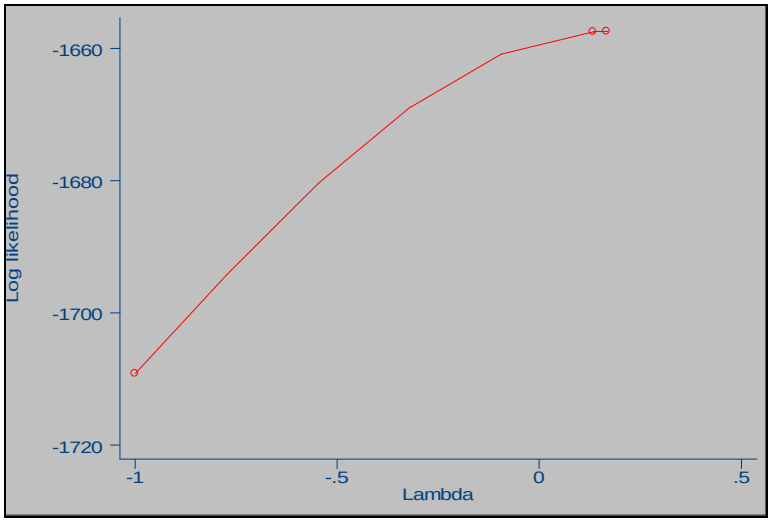
```
. swilk rstud

                  Shapiro-Wilk W test for normal data
    Variable |     Obs        W           V           z      Prob>z
-------------+--------------------------------------------------
       rstud |     314    0.93618     14.159       6.235   0.00000
```

a) What is your conclusion about the Normality assumption?

In order to correct the Normality problems we will use a transformed variable for plasma retinol levels. An easy way to find which transformation to use, is the general method of Box and Cox. This can be performed in STATA as follows (STATA will create a new variable named "newret" containing the transformed values):

```
.version 6
.  boxcox retplasm, lstart(-1) graph generate (newret)
(note: iterations performed using zero =.001)

   Iteration      Lambda         Zero     Variance        LL
   ------------------------------------------------------------
         0       -1.0000     89.67819   51344.1853  -1702.87019
         1        0.1327      2.56310   37070.2753  -1651.72960
         2        0.1676      0.00243   37062.7391  -1651.69768
         3        0.1676      0.00000    37062.742  -1651.69770
   ------------------------------------------------------------

Transform:  (retplasm^L-1)/L

                L       [95% Conf. Interval]     Log Likelihood
            ---------------------------------------------------
               0.1676      (not calculated)           -1651.6977

     Test:  L == -1      chi2(1) =   104.14    Pr>chi2 =  0.0000
            L ==  0      chi2(1) =     2.19    Pr>chi2 =  0.1387
            L ==  1      chi2(1) =    49.39    Pr>chi2 =  0.0000
. version 8
```

The graph produced plots the log likelihood versus λ. λ=0.1676 is the maximum likelihood estimate.



Now we can use the stepwise procedure again

```
.  xi: sw reg newret age i.sex (i.smokstat) quetelet (i.vituse) calories fat fi
> ber alcohol betadiet retdiet, pr(.1)
i.sex                 Isex_1-2     (naturally coded; Isex_2 omitted)
i.smokstat            Ismoks_1-3   (naturally coded; Ismoks_1 omitted)
i.vituse              Ivitus_1-3   (naturally coded; Ivitus_1 omitted)
                      begin with full model
p = 0.6699 >= 0.1000  removing retdiet
p = 0.6327 >= 0.1000  removing quetelet
p = 0.5945 >= 0.1000  removing Ivitus_2 Ivitus_3
p = 0.5018 >= 0.1000  removing betadiet
p = 0.2852 >= 0.1000  removing Isex_1
p = 0.1146 >= 0.1000  removing Ismoks_2 Ismoks_3

  Source |       SS       df       MS                  Number of obs =     314
---------+------------------------------               F(  5,   308) =    7.93
   Model |  34.2403962     5  6.84807924               Prob > F      =  0.0000
Residual |  265.811233   308  .863023485               R-squared     =  0.1141
---------+------------------------------               Adj R-squared =  0.0997
   Total |   300.05163   313  .958631405               Root MSE      =  .92899

------------------------------------------------------------------------------
  newret |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .0161012   .0038357      4.198   0.000       .0085538    .0236487
     fat |  -.0092023   .0042933     -2.143   0.033      -.0176502   -.0007544
calories |   .0004638   .0002721      1.704   0.089      -.0000716    .0009992
   fiber |  -.0235039   .0140948     -1.668   0.096      -.0512381    .0042303
 alcohol |   .0368435   .0115156      3.199   0.002       .0141843    .0595027
   _cons |    10.6249   .2747965     38.665   0.000       10.08418    11.16561
------------------------------------------------------------------------------
```

We generate the Student's residual again along with the Cook's d and the leverage in order to check for outliers and influential observations

```
. predict rstud, rstud

. predict d,cooksd

. predict h, hat
```

b) Check the normality assumption using the `swilk` command.

Now list residuals, leverages and Cook's d's for observation with residual more than 2 and leverage greater than 2p/n (2*6/314=0.0382). Then check the maximum Cook's d.

```
. list rstud d h if abs(rstud)>2.0 | h>.0382

         rstud          d          h
  1. -2.880156   .0283224   .0205401
  2. -3.392216   .0174023   .0092961
  .
  .
  313.  3.294111   .0123143   .0069778
  314.  3.587223   .0196182   .0094104

(37 cases)
. list rstud d h if abs(rstud)>2.0 & h>.0382
( 0 cases)
```

```
. summarize d

Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
       d |     314    .0031528    .0058947   9.35e-11    .0390723
```

c) There are no observations with Cook's distance above 1, although there are several points with large residuals or leverage. What do you think about the "residual>2" criterion?