GENERALIZED LINEAR MODELS:Model sellection

Γιώτα Τουλούμη

Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας

Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής

Ιατρική Σχολή Πανεπιστημίου Αθήνας

gtouloum@med.uoa.gr

Model selection

To motivate model selection in the generalized linear model, I present the mechanics of model selection in the linear model.

Consider the process of starting with a "full" model in the sense that it is a model containing all variables that we are willing to consider. Then the criterion of removing a variable is based on an F test as follows (here we consider p variables plus the intercept in all models):

$$\frac{SSE(X_{p_1}) - SSE(X_{p_2})}{SSE(X_{p_2}) \Big/ (n - p_2 - 1)} \sim F_{1, n-p_2-1}$$

where $SSE(X_{p_2})$ and $SSE(X_{p_1})$ are the residual sum of squares of the full model and the sub-model respectively

Example: Plasma retinol levels (continued). The output from the full model is as follows:

```
. 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
------------------------------------------------------------------------------
```

While the output from the model excluding cholesterol levels is,

```
. 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
```

The criterion for removing cholesterol level from consideration is

$$\frac{SSE(X_{p_1}) - SSE(X_{p_2})}{SSE(X_{p_2})/(n-p_2-1)} = \frac{11744522.4 - 11723197.4}{11723197.4/301} = 0.5475$$

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

This can also be given by using the `test` command after regression on the full model as follows:

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

. test chol

 ( 1)  chol = 0.0

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

and is equivalent to the *t* test listed in the output of the full model above (recall that an *F* test with 1 degree of freedom in the numerator is equal to the square of the *t* test with equal degrees of freedom as in the denominator of the *F* test).

Model selection in the GLM

A similar concept as the residual sums of squares in the GLM is the *deviance*. In addition, the log-likelihood can be used in the derivation of likelihood-ratio tests. We consider these two concepts here.

The likelihood ratio λ is the fraction of the maximized likelihood of the sub-model and the full model respectively. For large samples, $-2\log\lambda \sim \chi^2_\nu$ where ν is the difference in the dimension of the two models. For two models that are different by a single variable, $\nu=1$ of course.

In general, the likelihood-ratio criterion is $-2\log\lambda = \dfrac{D(X_{p_1}) - D(X_{p_2})}{\varphi}$, where φ is a scale parameter, and $p_1 < p_2$. In particular, in the linear model this is is,

$$-2\log\lambda = \frac{SSE(X_{p_1}) - SSE(X_{p_2})}{SSE(X_{p_2}) \Big/ (n - p_2 - 1)}$$

Example: Plasma retinol levels (continued)

In our example, we can derive the likelihood-ratio test as follows:

```
. xi: glm retplasm i.sex age i.smokstat i.vituse  quetelet calories fat fiber alcohol  cho
> l
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 param     =   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              =  11723122.68

------------------------------------------------------------------------------
    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
```

Example: Plasma retinol levels (continued)

```
. xi: glm retplasm i.sex age i.smokstat i.vituse  quetelet 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)

Iteration 0:   log likelihood = -2098.6789

Generalized linear models                         No. of obs      =        314
Optimization     : ML: Newton-Raphson             Residual df     =        302
                                                  Scale param     =   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
------------------------------------------------------------------------------
```

The likelihood-ratio test can be constructed as follows:

$$-2\log\lambda = \frac{SSE(X_{p_1}) - SSE(X_{p_2})}{SSE(X_{p+1})/(n-p-1)} = \frac{11744522.37 - 11723122.68}{11723122.68/301} = 0.5477$$

Its asymptotic (long-term) distribution is a chi-square with one degree of freedom.

```
. di chiprob(1,(11744453.38-11723122.68)/((11723122.68)/(301)))
.45926647
```

which is similar to the results of the F test previously. Notice that we get the same results if we subgract the maximized log-likelihoods as follows:

$$-2\log\lambda = -2\left[-2098.67891-(2098.39358)\right] = 0.57066$$

with asymptotic distribution that is also chi-square with one degree of freedom.

```
. di chiprob(1,-2*(2098.39358-2098.67891))
.44999677
```

Wald tests

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

```
. quietly xi: glm retplasm i.sex age i.smokstat i.vituse  quetelet calories fat fiber alco
> hol  chol
```

In STATA 7.0, this is given by

```
. test chol

 ( 1)  [retplasm]chol = 0.0

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

In STATA 6.0, we can derive the chi-square (Wald) test as follows:

```
. di chiprob(1,( -.0775529/.1048078)^2)
.4593282
```

Finally, we show here the model-selection for the complete problem.

```
. xi: sw glm retplasm  i.sex (i.smokstat ) (i.vituse) age quetelet calories fat
>  fiber alcohol  chol, 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_3 omitted)
                      begin with full model
p = 0.4599 >= 0.1000  removing chol
p = 0.4231 >= 0.1000  removing quetelet
p = 0.4163 >= 0.1000  removing Ivitus_1 Ivitus_2
p = 0.1572 >= 0.1000  removing Ismoks_2 Ismoks_3
p = 0.1806 >= 0.1000  removing fiber
p = 0.5284 >= 0.1000  removing calories

Residual df  =       309                              No. of obs =       314
Pearson X2   =  1.21e+07                              Deviance   =  1.21e+07
Dispersion   =   39055.6                              Dispersion =   39055.6

Gaussian (normal) distribution, identity link
------------------------------------------------------------------------------
retplasm |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Isex_1 |     74.055   36.44476      2.032   0.043       2.343714    145.7663
     fat |  -.6188433   .3501419     -1.767   0.078      -1.307807    .0701208
 alcohol |   8.724091   2.340494      3.727   0.000        4.11877    13.32941
     age |   2.389427   .8229901      2.903   0.004       .7700534    4.008801
   _cons |   498.7073   54.22216      9.197   0.000       392.0159    605.3986
------------------------------------------------------------------------------
(Model is ordinary regression, use regress instead)
```

Pearson Residuals

The Pearson residuals are defined as

$$r_{i,p} = \frac{y_i - \hat{\mu}_i}{\left[V\left(\hat{\mu}_i\right)\right]^{1/2}}$$

and it is the raw residual scaled by the estimated standard deviation of Y. The name is taken from the fact that for the Poisson distribution the Pearson residual is just the signed square root of the component of the Pearson X^2 goodness-of-fit statistic, i.e.

$$\sum_{i=1}^{n} r_{i,p}^2 = X^2$$

A disadvantage of the Pearson residual is that the distribution of $r_{i,p}$ for non-normal distributions is markedly skewed, and it may fail to have properties similar to those of a normal-theory residual.

Deviance Residuals

If the deviance is used as a measure of discrepancy of a generalized linear model, then each unit contributes a quantity d_i to that measure, so

$$\sum d_i = D$$

Thus, if we define

$$r_{i,D} = sign(y_i - \mu_i)\ \sqrt{d_i}$$

we have a quantity that increases with $y_i - \boldsymbol{\mu}_i$ and for which $\sum r_{i,D}^2 = D$.

Residuals- Linear regression

Recall that variance of the true residuals is assumed to be constand. The variance of the fitted (observed) residuals is NOT constant, since there is variance in estimation of the line and of the expected values. Therefore, for model checking we need to standardize the observed residuals.

Lets explore it in normal regression

$$E(Y|X=x)=X\beta$$

design matrix

$$H=X\left(X^TX\right)^{-1}X^T$$ hat matrix

$$h_{ii}=h_i=x_i^T\left(X^TX\right)^{-1}x_i$$ i^{th} leverage

$$e=y-\hat{y}$$ residuals and

$$Var(e)=(I-H)\sigma^2$$

Residuals- Linear regression (continue)

standardized residual $r_i = \dfrac{e_i}{S\sqrt{1-h_i}}$

sample variance

here e_i, S are not independent since e_i enters in the calculation of S.

studentized residuals $r_i^* = \dfrac{e_i}{S_{(i)}\sqrt{1-h_i}}$

sample variance with i^{th} observation omitted

here numerator & denominator are independent

The distribution of $r_i^* \sim t_{n-p-1}$

The i^{th} leverage is large if $h_i \geq 2p'/n$ where $p' =$ total # of covariates in the model including intercept, $n =$ total # of observations.

Standardized residuals in GLMs

The key quantities for GLM diagnostics are:

- residuals → Pearson
- residuals → Deviance

The general definition of standardized residuals is:

General definition

$$r'_p = \frac{y - \hat{\mu}}{\sqrt{\hat{\phi} V(\hat{\mu})(1-h)}}$$

$$r'_D = \frac{r_D}{\sqrt{\hat{\phi}(1-h)}}$$

leverage

leverage

(standardized deviance residual)

Model checking

The predicted values and the residuals from the optimal model (the one including gender, fat and alcohol intake and age) are produced by STATA commands as follows:

```
. quietly reg retplasm sex fat alcohol age

. predict yhat
(option xb assumed; fitted values)

. predict r, resid

. predict rstan, rstand

. predict rstud, rstud
```

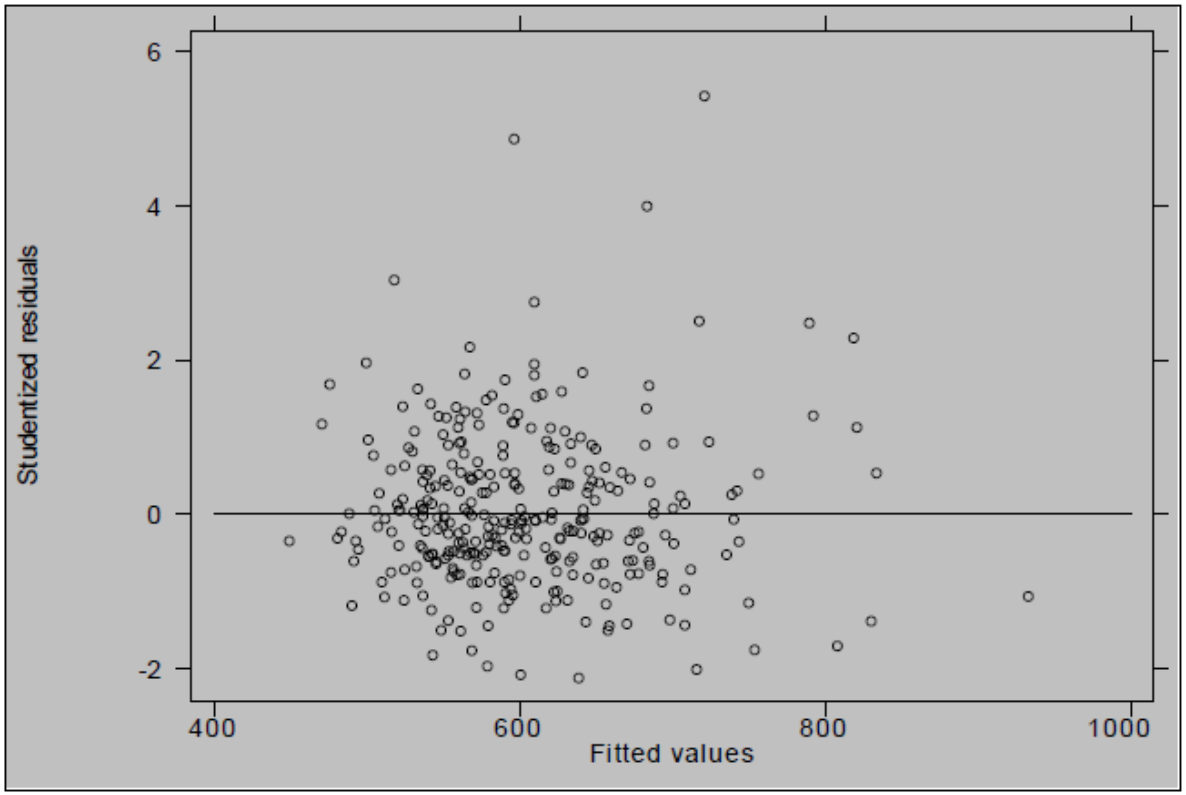
Model checking: residuals

The assumptions of the model that must be checked are independence, normality and homoskedasticity. We usually work with the standardized residuals $r_{std,i} = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ (produced with the option `rstan`) or the studentized residuals $r_{stud,i} = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$ (with option `rstud`), where $\hat{\sigma}$ is an estimate of the standard deviation derived with all the observations, while and $\hat{\sigma}_{(i)}$ is the estimate with the i^{th} observation missing. On the other hand, h_{ii} is the i^{th} diagonal element of the *hat* matrix (recall that in regression $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ is the "hat" matrix). The leverage points are a measure of *distance (outlier, potential influential point)*. We use the Cook's distances as a combined measure of influence and distance since they are $D_i = r_{std,i}^2 \frac{h_{ii}}{(1-h_{ii})}$.

Homoskedasticity

This refers to the homogeneity of variance. We can see what the stud. residuals look like as follows:

```
. graph r yhat,yline(0) xlab ylab border
```

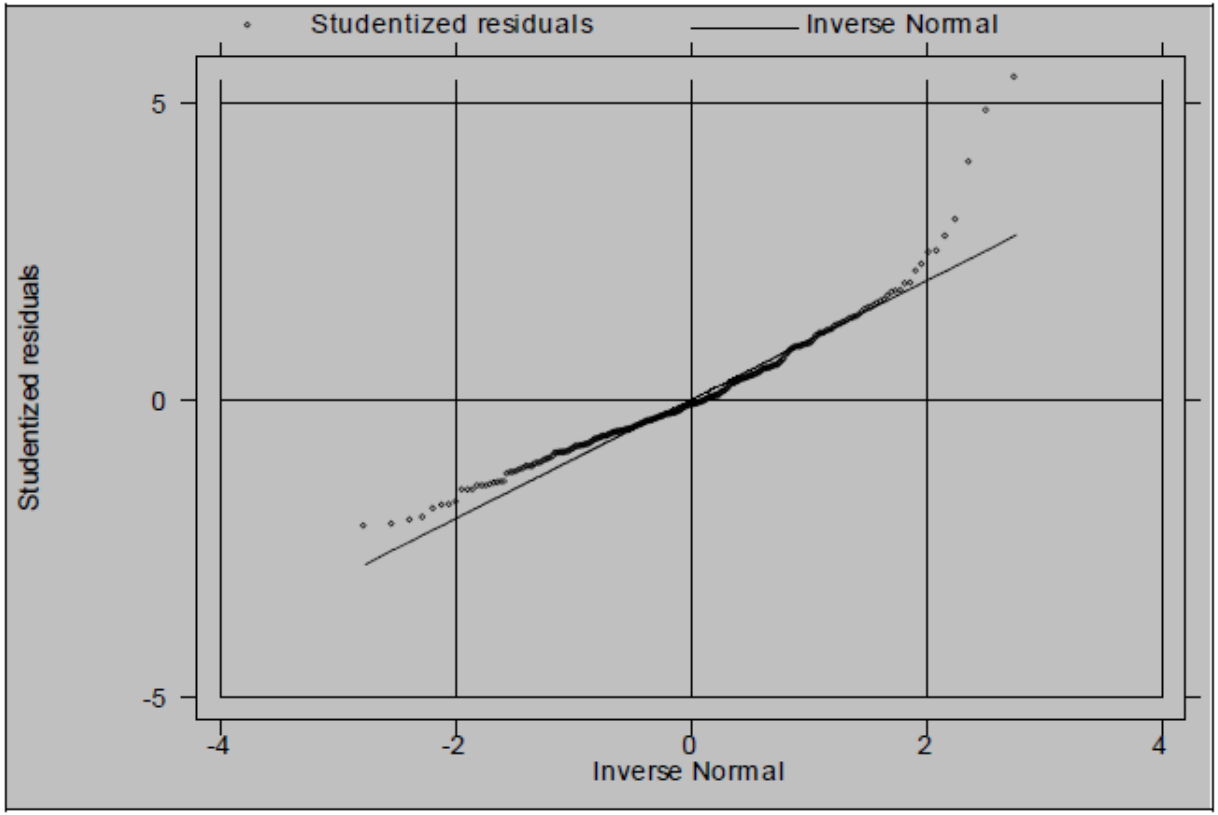


We see that there is no obvious problem with lack of homoskedasticity in these data.

Normality

The next assumption of the general linear model is that of normality of the residuals.

This can be checked using the `qnorm` command in STATA as follows:



Q-Q plots

These are plots that compare the distribution of a variable to a known distribution. They can be used alternatively to compare the distributions of two variables. In general, if the distributions are approximately equal the points on the graph should lie on a straight line.

In the plot above, we see that there are problems with the distribution of retinol levels at the "tails" which are shorter for small values and "fatter" for larger values.

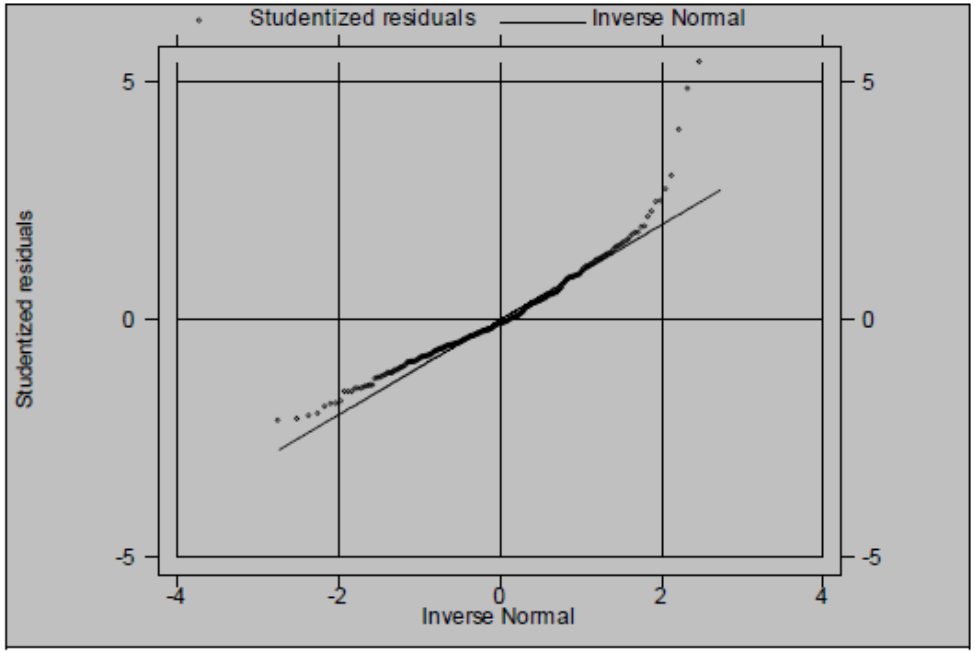
This Q-Q plot can be produced manually following these steps:

- Sort the residuals from smaller to largest (we are still working with studentized residuals)
- Imagining that the residuals have a normal distribution then their ranked values should be close to standard normal distribution percentiles, that is, $z_{(i)} = \Phi^{-1}\left(\frac{i}{n+1}\right), i = 1,\ldots,n$ (this is actually the way STATA produces a q-q plot).

Q-Q plot (continued)

To create the q-q plot above manually we proceed as follows:

```
. sort rstud
. gen zi=invnorm(_n/(_N+1))
. label var zi "Inverse Normal"
. graph rstud zi zi, xlab ylab c(.l) s(oi) rlab yline xline
```



The Shapiro-Wilks test of normality

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
```

The test p value is 0.000<0.05 which means that the normality assumption is not fulfilled.

Box-Cox transformations

In order to find which transformation to use, a general method is that of Box and Cox. The general Box-Cox transformation is as follows:

$$y^* = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & y \neq 0 \\ \log(y), & y = 0 \end{cases}$$

Several possible choices of λ are tried. The best choice is given through a likelihood criterion.

Some usual transformations are given as follows:

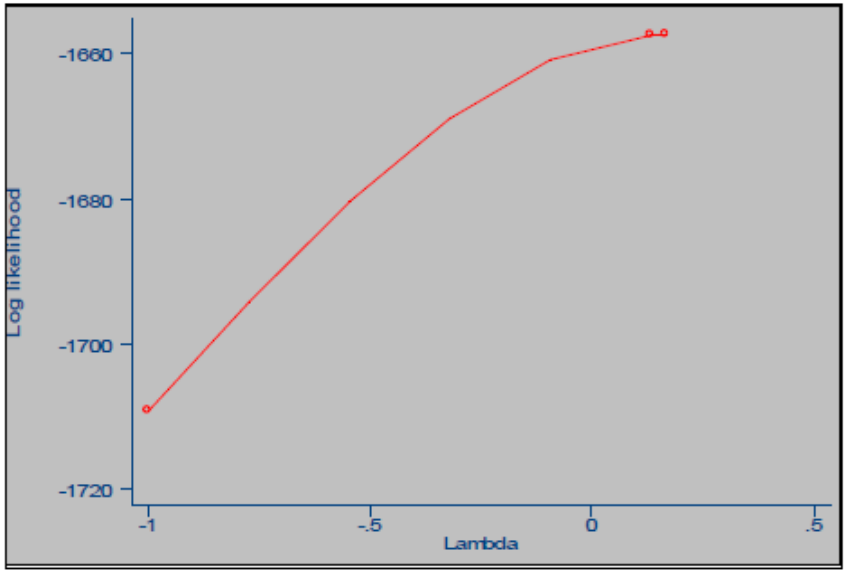
- λ=-1 Inverse transformation
- λ=1 No transformation is necessary
- λ=0.5 Square-root transformation
- λ=0 Logarithmic transformation

Box-cox transformation

To implement the Box-Cox technique in STATA we proceed as follows:

```
.  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
```

Box-cox transformation (continued)



A value of zero for lamda is not unreasonable, suggesting a logarithmic transformation of (`retplasm`). The new variable `newret` contains the transformed values of `retplasm` (with lamda = 0.1676).

Further model checking

Running the model `newret` with as the dependent variable we have:

```
.  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
```

Ερμηνεία συντελεστών

$$E(\ln($$

Checks for outliers and influential observations

We produce residuals, *leverage* values and Cook's distances as follows:

```
. predict rstud, rstud

. predict d,cooksd

. predict h, hat
```

A studentized residual greater than 2 in absolute value, a leverage greater than $2p/n$=0.0382, where p is the number of predictors plus the intercept, and a Cook's distance of 1 or higher are indicative of an outlier or of excessive influence, or both respectively.

```
. 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)
```

Model checking (continued)

To `summarize` the Cook's distances we proceed as follows:

```
. summarize d

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

There are no obsrervations with Cook's distance above 1, although there are several points with large residuals or leverage. However, the number of points that we are testing for large residuals is so large, that the criterion of 2.0 or higher is probably very liberal (as 314 repeated tests are being conducted!). Thus, the fit is probably acceptable.