Epidemiology/Biostats VHM 812/802 Course
Winter 2015, Atlantic Veterinary College, PEI
Javier Sanchez

Lecture 2b: Linear Regression Diagnostics

Index **Page**

- Exercises – Monday Lab
  - Linear regression exercise 2
  - Linear regression exercise 3

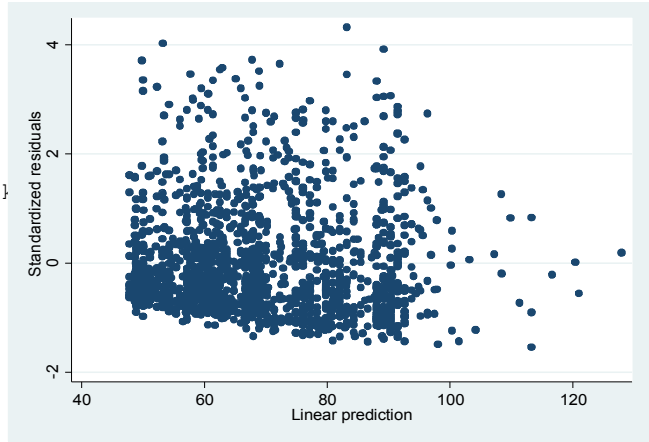
Example: WPC model (daisy2red.dta)

- VER example 14.12
  - outcome: wpc (interval from waiting period to conception)
  - predictors: parity, twin, dyst, rp, vag_disch, herd_size and herd_size2, calving_date (in months)
    - also interactions: rp*vag_disch
- final (candidate) model

```
. reg wpc hs100_ctr hs100_ctr_sq parity1 aut_calv twin dyst rp##vag_disch, vsquish

      Source |       SS       df       MS              Number of obs =    1574
-------------+------------------------------           F(  9,  1564) =   13.22
       Model |  296062.694     9  32895.8549           Prob > F      =  0.0000
    Residual |  3892027.86  1564  2488.50886           R-squared     =  0.0707
-------------+------------------------------           Adj R-squared =  0.0653
       Total |  4188090.56  1573  2662.48605           Root MSE      =  49.885

------------------------------------------------------------------------------
         wpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   hs100_ctr |   19.85708   2.163397     9.18   0.000     15.61361    24.10054
hs100_ctr_sq |   11.13827   3.111145     3.58   0.000     5.035817    17.24073
     parity1 |    1.13721   .8583103     1.32   0.185    -.5463501     2.82077
    aut_calv |  -8.263839   2.537751    -3.26   0.001    -13.24159   -3.286086
        twin |   20.68314   9.845165     2.10   0.036      1.37203    39.99425
        dyst |   11.70041   5.462576     2.14   0.032      .985666    22.41516
          rp |
         yes |    5.98687   4.811976     1.24   0.214    -3.451734    15.42547
   vag_disch |
         yes |   1.228196   7.161395     0.17   0.864    -12.81875    15.27514
rp#vag_disch |
    yes#yes  |   22.85194   12.51605     1.83   0.068    -1.698056    47.40194
       _cons |   64.33029   2.634114    24.42   0.000     59.16352    69.49705
------------------------------------------------------------------------------
```

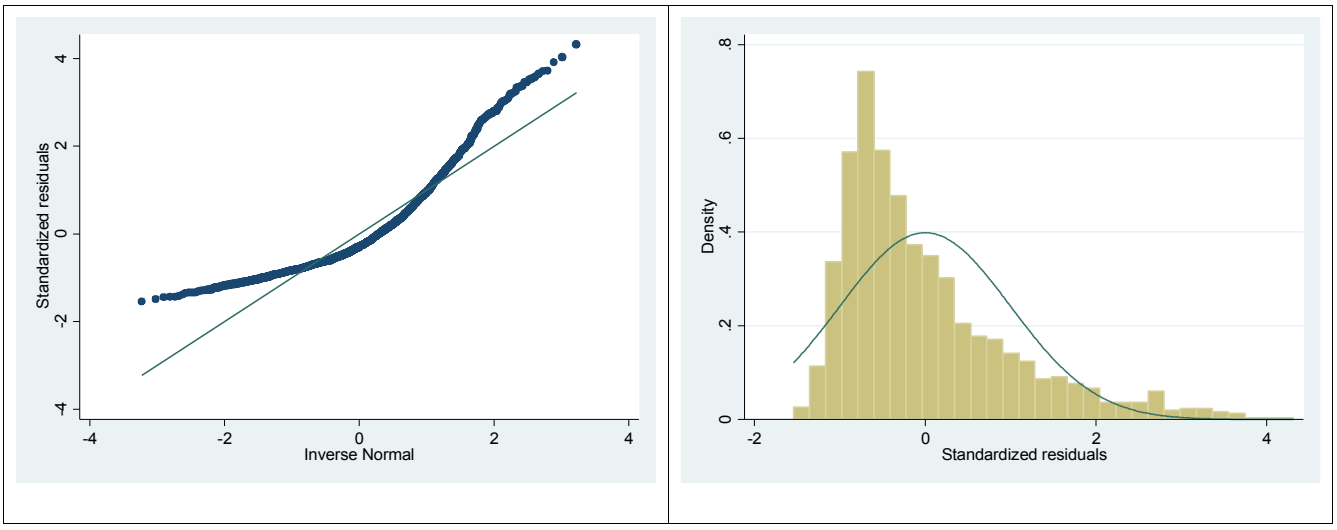
Evaluating major assumptions

- homoscedasticity




```
. hettest

Breusch-Pagan / Cook-Weisberg test for heterosk
         Ho: Constant variance
         Variables: fitted values of wpc

         chi2(1)      =    20.58
         Prob > chi2  =   0.0000

. imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df       p
---------------------+-----------------------------
  Heteroskedasticity |      74.11     44    0.0030
            Skewness |     143.84      9    0.0000
            Kurtosis |      33.27      1    0.0000
---------------------+-----------------------------
               Total |     251.22     54    0.0000
---------------------------------------------------
```

- normality




```
. swilk  stdres

                    Shapiro-Wilk W test for normal data

    Variable |     Obs       W           V         z       Prob>z
-------------+--------------------------------------------------
      stdres |    1574    0.87871     115.660    11.977    0.00000
```

Transformation of Y (Henrik's notes)

- Box-cox
- interpretation

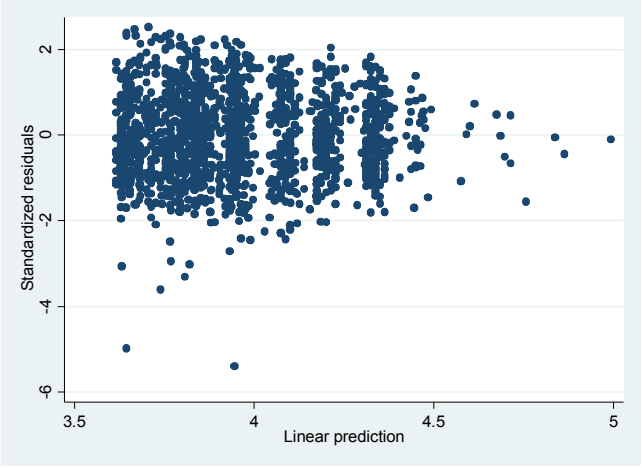
Re-assessing assumptions

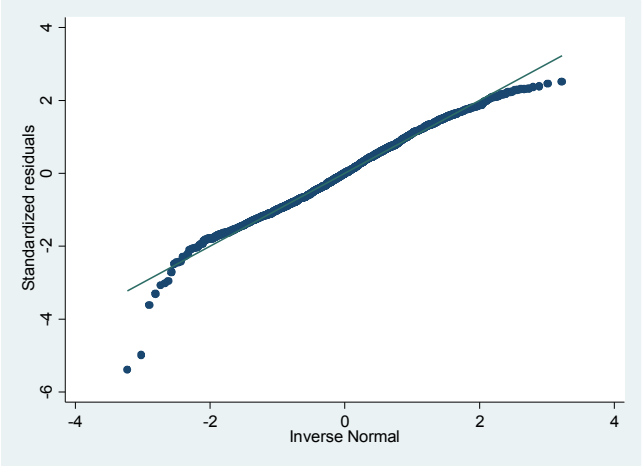
- Log-transformed model

```
. boxcox wpc aut_calv hsize hsize2 parity1 twin dyst rp vag_disch
...output omitted...

                                                Number of obs   =       1574
                                                LR chi2(8)      =     148.83
Log likelihood = -7960.5368                     Prob > chi2     =      0.000

------------------------------------------------------------------------------
         wpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   .1099104   .0271003     4.06   0.000     .0567947    .1630261
```

  - ★ value of λ close to 0 -> log transformation

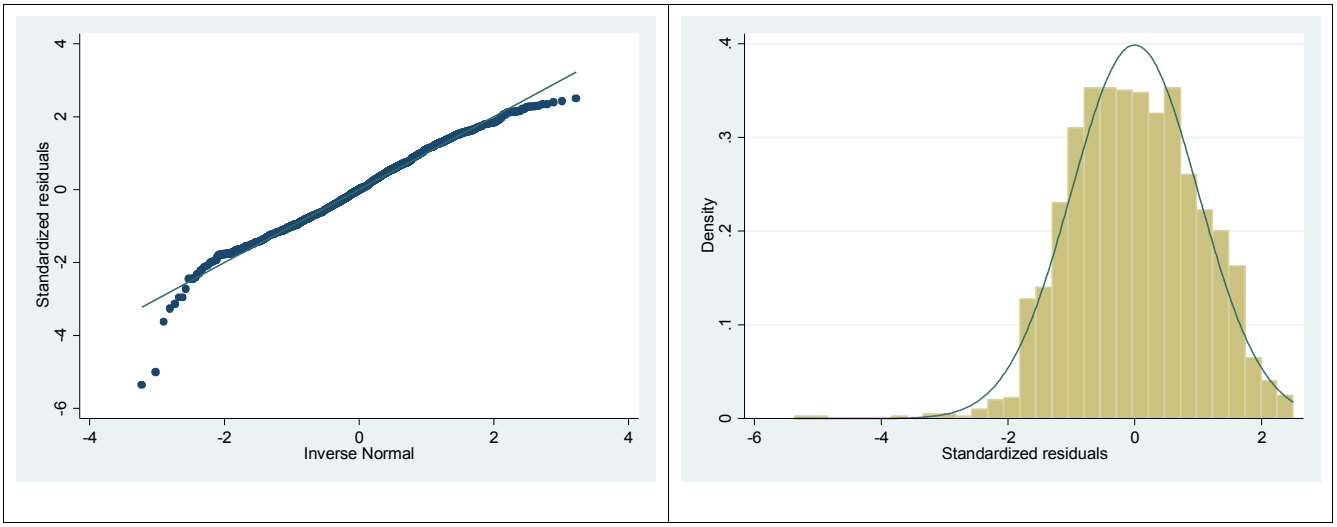
- homoscedasticity





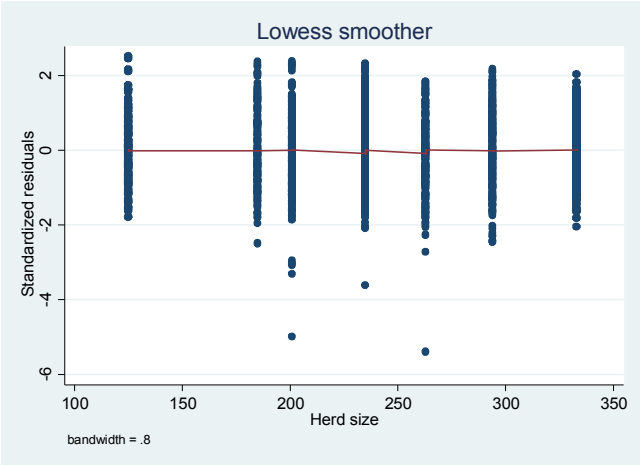
  - ★ tests provide contradicting results
    - imtest – constant variance
    - however BPCW test highly significant

- normality

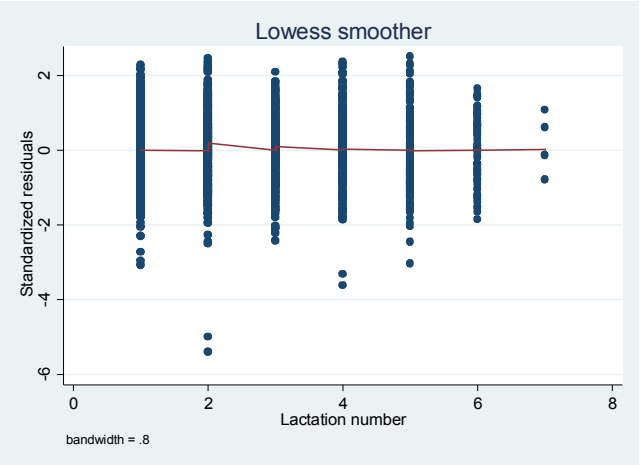


- linearity
  - ★ continuous predictors only
  - ★ standardized residuals vs predictor

Herd size



Parity



  - ★ also pattern can be evaluated by looking at the residuals from a model without the predictor of interest

Evaluating individual observations

Outliers (lecture L1a – pages 8/9)

- large residuals
- standardized residuals
  - expect 5% > 2 and < -2
  - expect 1% > 3 and < -3
- deletion residuals
  - t-test - prob = 2*n*t(DFE-1, d_i)
- wpc log-transformed model – residuals
  - standardized residuals
    - expected 5% = 79 (observed = 50)
    - expected 1% = 14 (observed = 6)
    - fewer values than expected
  - deletion residuals
    - outlier cutoff value = 4.17
    - two observations >= this cutoff point
      - indication that these are outliers
        - cows with wpc = 1
    - delete and refit the model (later)
      - only for diagnostics purposes

Leverage

- x-outliers
  - depend on X values only
    - not affected by relationship with y
  - expresses whether x_i is outlying in distribution of x's
    - so potentially influential
  - less meaningful for categorical predictors
    - depends on the distribution of each category

- cut-off values
  - concerns if $h_i \geq 2(k+1)/n$
    - or $\geq 3(k+1)/n$
    - k=# of predictors, n=# obs.
    - should also look for unusual observations independently of these cutpoints

- Stata – *predict* varname, lev<u>erage</u>
  - eg. predict lev_lnwpc, lev

- ln_wpc full model

```
. summ lev

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         lev |      1574    .0063532    .0083435   .0016636   .0642237

. display "leverage cutoff:  " 2*nparam/nobs
leverage cutoff:  .01270648

. display "conservative leverage cutoff:  " 3*nparam/nobs
conservative leverage cutoff:  .01905972

. count if lev>=.01905972 /* many (n=108) high leverage values */


. list wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres lev in 1/10, clean noobs

    wpc   wpc_ln     fit   aut_calv   herd_s~e   parity1   twin   dyst    rp   vag_di~h   stdres     lev
     76    4.331   4.690        yes        185         4    yes     no   yes        yes   -0.507   0.065
    137    4.920   4.989        yes        294         2    yes     no   yes        yes   -0.097   0.064
     94    4.543   4.859        yes        263         3    yes     no   yes        yes   -0.445   0.064
     45    3.807   4.612         no        201         4    yes     no   yes        yes   -1.135   0.063
     53    3.970   4.251         no        263         5    yes     no    no        yes   -0.395   0.059
    110    4.700   4.188         no        263         0    yes     no    no        yes    0.720   0.057
     32    3.466   4.286         no        201         1    yes    yes   yes         no   -1.150   0.052
     99    4.595   4.575        yes        294         1    yes    yes    no         no    0.029   0.051
     23    3.135   4.143         no        185         0    yes    yes    no         no   -1.409   0.050
    106    4.663   4.600         no        294         0     no    yes   yes        yes    0.088   0.046

. tab3way twin rp dyst

------------------------------------
           | Dystocia at calving and
           |   Retained placenta at
           |          calving
Twins      | --- no ---     --- yes --
born       |   no    yes       no    yes
-----------+------------------------
       no  | 1332    124       76     15
      yes  |   15      9        2      1
------------------------------------
```

  - large leverage values for less common combination of categorical predictors

Influence diagnostics: Cook's distance and DFITS

- Cook's distance and DFITS

Cooks D

$$D_i = \frac{r_{si}^2}{(k+1)} * \frac{h_i}{(1-h_i)}$$

DFITS

$$DFITS_i = r_{ti}\sqrt{\frac{h_i}{(1-h_i)}}$$

  - measure influence of an observation and have two interpretations:
    - effect of deleting observation "i" on the predictions
    - effect of outlier information about x's and y's
  - D_i is on the squared residual scale and based on standardized residuals
  - $DFITS_i$ is on absolute (signed) residual scale and are based on deletion residuals

- cut-off values
  - concern if either is >1 (rare) or if...
    - $D_i \geq 1$ or $\geq 4/n$
    - $DFITS_i$ outside $\pm 2*\sqrt{(p/n)}$ (if n ≥ 120)
      - n < 120 only values beyond +/- 1

- Stata: *predict* varname, cooksd/dfits

- ln_wpc model
  - Cook's

```
. summ cook

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        cook |      1574    .0006351    .0013679   5.57e-10   .0174304

.display "Cook's D cutoff:   " 4/nobs
   Cook's D cutoff:   .0025413

. count if cook>=.0025413 /* many (n=94) high Cook's D values */
   94
```

  - DFITS

```
. summ dfit

   Variable |       Obs        Mean    Std. Dev.       Min        Max
------------+--------------------------------------------------------
       dfit |      1574    .0003931    .0797594  -.4179343   .3664328


. display "DFITS cutoff:   " 2*sqrt(nparam/nobs)*(nobs>=120)+1*(nobs<120)
      DFITS cutoff:   .15941443

. count if abs(dfit)>=.15941443 /* many (n=92) high DFITS values */
   92
```

  - first lactation cows with reproductive diseases have highest Cook's and DFITS values

Delta-beta (or DFBETA)

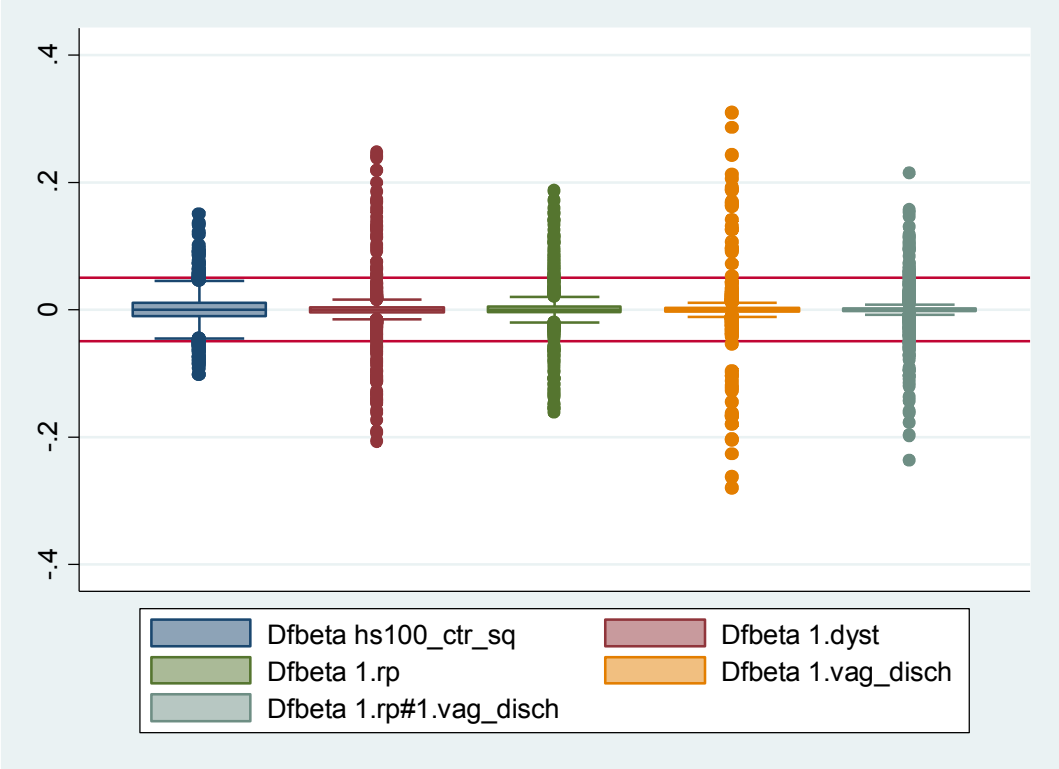
- influence of an observation on a specific regression coefficient (β)
- for obs. "i" and predictor x_j
  - $DFBETA_{ij} = \frac{(\beta_j - \beta_{j(i)})}{SE_j}$
  - where $\beta_{j(i)}$ is estimate of β without obs. "i"
- measures change in β en terms of SE's
  - how many SE's does β_j increase if "i" is dropped?

- cut-off values
  - $DFBETA_{ij}$ > 0 (<0) ~ obs. "i" pulls β_j up (down)
    - extreme values of $DFBETA_{ij}$ are notable
    - most meaningful for continuous predictors
  - concern if outside $\pm 2/\sqrt{(n)}$ (for n ≥ 120)
    - n < 120 only values beyond +/- 1

- Stata
  - *predict* varname, dfbeta(var in model)
    - eg. predict dfbdyst, dfbeta(dyst)

- ln_wpc model

```
. display "DFBETA cutoff:   " 2/sqrt(nobs)*(nobs>=120)+1*(nobs<120)
DFBETA cutoff:   .05041127
```



- explore values by browsing/listing
  - large values are cases of the correspondent predictor (eg. cows with dystocia)
  - linear effect of herd size similar to quadratic term
    - need to be aware of collinearity

What to do with Outliers/Influential observations

- verify that point are not errors (eg. data entry errors)
- may be very informative
  - always examine observations and values of the predictors
- fit models with and without the point(s)

- keeping them in the analysis
  - potential for biased estimates
  - usually leads to larger SE and reduced power (is therefore conservative)
  - always report them
- eliminating them from the analysis
  - **should be always reported and justified**
  - narrows scope of inference from the study
  - may creates an unrealistically good model

Other transformations

- log-transformed model
  - some indication of non-constant variance
  - residuals still not normally-distributed
    - two severe outliers

```
. l wpc wpc_ln  fit herd_size parity dyst rp vag_disch if wpc==1, clean compress
noobs

    wpc   wpc~n        fit   her~e   par~y   dyst   rp   vag~h
      1       0   3.645548     201       2     no   no      no
      1       0   3.945861     263       2     no   no      no

. l wpc wpc_ln  delres lev dfit cook dfb_dyst dfb_rp dfb_vd dfb_int dfb_hs if wpc==1, clean
compress noobs

    wpc   wpc~n   delres     lev     dfit    cook   df~st   dfb~p   dfb~d   dfb~nt   dfb_hs
      1       0   -5.028   0.002   -0.243   0.006   0.051   0.037   0.028   -0.021    0.098
      1       0   -5.449   0.002   -0.243   0.006   0.044   0.053   0.042   -0.026   -0.016
```

  - comparison without outliers

```
. estimate table ln ln_noout
-------------------------------------
    Variable |     ln        ln_noout
-------------+-----------------------
   hs100_ctr |  .34365799    .33916323
hs100_ctr_sq |  .21187276    .20269649
     parity1 |  .01298436     .0113333
    aut_calv | -.13716214   -.13695147
        twin |  .39274261     .3894441
        dyst |   .1109405    .10338895
             |
          rp |
         yes |  .11231661    .10603408
             |
   vag_disch |
         yes | -.02601346   -.03328151
             |
rp#vag_disch |
    yes#yes  |  .41369992    .42230092
             |
       _cons |  3.8885867    3.9010244
-------------------------------------
```

  - repeat but without influential obs.
  - advantage log-transformed model
    - interpretation of estimates still possible

- square root transformation
  - example VER – constant variance ->ok
  - residuals still not normally-distributed
  - deletion residuals show no indication of outlying observations
  - interpretation estimates not possible
    - obtain predicted values
    - parity=3 aut_calv=1 hsize=0 hsize2=0 twin=0

	Sqrt transformation			log-transformation		
rp#vag_disch#dyst	**est**	**lo_ci**	**up_ci**	**est**	**lo_ci**	**up_ci**
no#no#no	58.267	54.094	62.594	50.127	46.598	53.922
no#no#yes	66.840	63.016	70.776	56.008	52.629	59.604
no#yes#no	58.064	52.980	63.382	48.840	44.659	53.412
no#yes#yes	66.622	65.094	68.168	54.570	53.239	55.934
yes#no#no	64.364	48.202	82.859	56.085	42.254	74.444
yes#no#yes	73.359	63.454	83.983	62.666	53.554	73.327
yes#yes#no	90.259	90.259	90.259	82.645	82.645	82.645
yes#yes#yes	100.857	90.533	111.737	92.342	80.406	106.049

*ci estimated from t-distribution (Stata uses a different estimation procedure)

  - sqrt model – better than log-transformed
    - no outliers (eg. Deletion res. range: -2.42; 3.28)
  - influential diagnostics
    - similar to log-transformed model
    - more in VER.