Epidemiology/Biostats VHM 812/802 Course
Winter 2014, Atlantic Veterinary College, PEI
Javier Sanchez

Lecture 2b: Linear Regression Diagnostics

Index **Page**

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

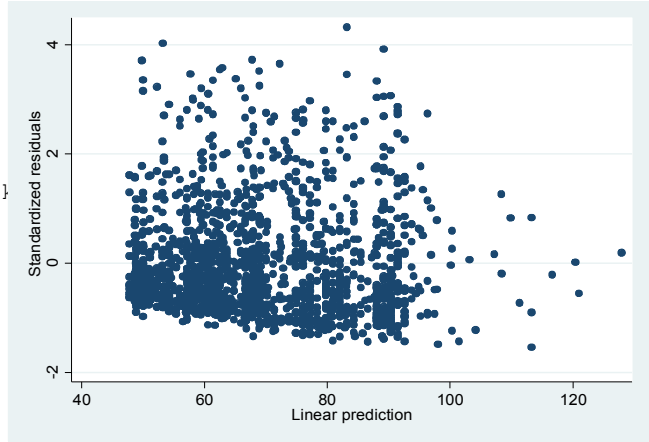
Example: WPC model (daisy2red.dta)

- 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 hsize hsize2 parity1 aut_calv twin dyst i.rp##vag_disch

      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]
-------------+----------------------------------------------------------------
       hsize |   19.85708   2.163397     9.18   0.000     15.61361    24.10054
      hsize2 |   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
------------------------------------------------------------------------------
```

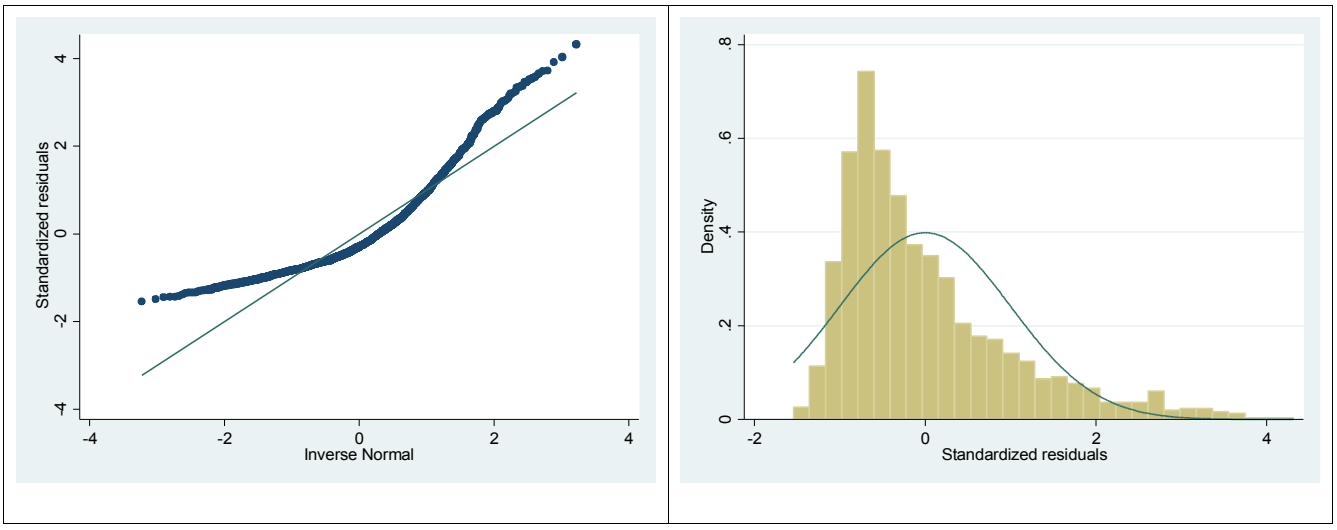
Evaluating major assumptions

- 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

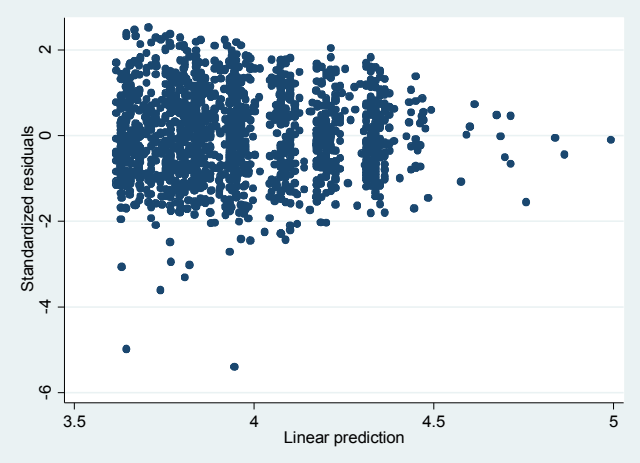


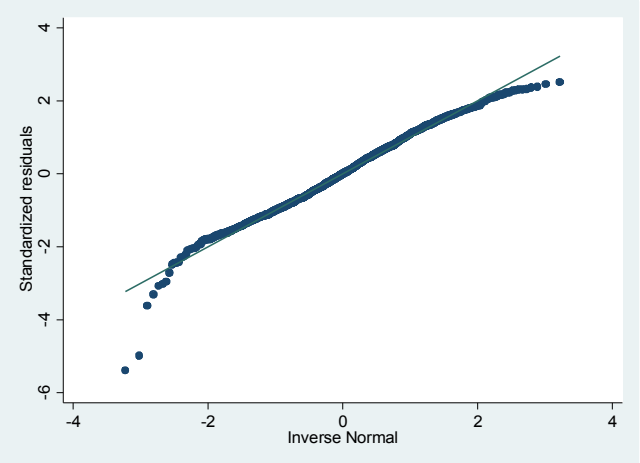

```
. swilk  stdres

                    Shapiro-Wilk W test for normal data

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

- transforming Y
  - Box-cox
    - y^{λ} ->choose value that makes residuals distribution close to Normal – with homogeneous variance
      - λ = 1 then Y* = Y (no transformation)
      - λ = 1/2 then Y* = $\sqrt{(Y)}$
      - λ = 0 then Y* = ln(Y)
      - λ = -1 then Y* = 1/Y

```
. 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

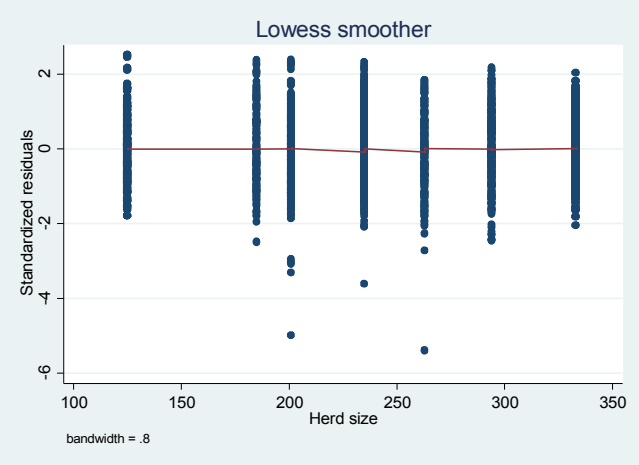




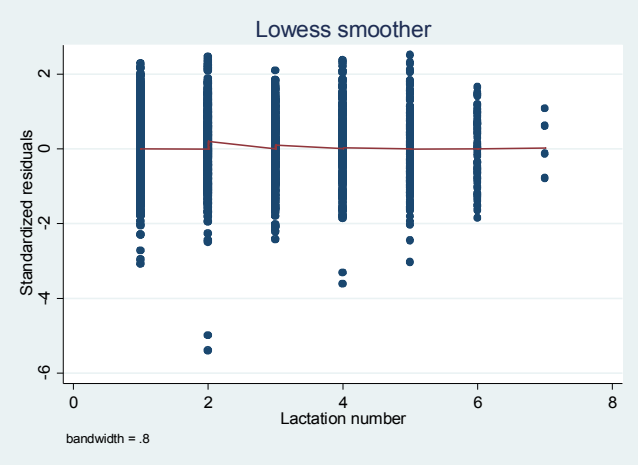
  - homoscedasticity test might suggest constant variance (imtest)
  - however BPCW test highly significant

- 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

Interpreting a transformed model

- problems with interpretation
  - results need to be backtransformed to original scale
  - statistical significance OK
- prediction
  - backtransformed means median wpc values
    - CIs can also be backtransformed
  - need to fix the values of the other predictors, unless their distribution is representative of target population
- estimates (only for log-transform)
  - $\hat{\beta}_{dyst} = 0.111 (\text{CI}: -0.046, 0.268) \text{--> } \exp(\hat{\beta}_{dyst}) = \exp(0.111) = 1.12$
  - $\text{CI}: (e^{-0.046,} e^{0.268}) \text{--> } = (0.955, 1.307)$
  - dystocia increases wpc by a factor of 1.12 or 12%
    - CI goes from a reduction of ~ 4% to an increase of ~31% in wpc
    - CI includes 1 -> non significant

Evaluating individual observations

Outliers (lecture 1al – 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 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

- simple linear regression

$$h_i = \frac{1}{n} + \frac{(X_{1i} - \bar{X}_1)^2}{SSX_i}$$

  - in general can be defined by:

$$Var(\hat{r}_i) = \sigma^2 * (1 - h_i)$$

- depend on X values only
- expresses whether x_i is outlying in distribution of x's
  - so potentially influential
- 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

- not affected by relationship with y
- less meaningful for categorical predictors
  - depends on the distribution of each category

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

. 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   4.330733   4.698747          0        185         4    yes     no   yes        yes   -.5200757   .0642237
     45   3.806663   4.577247          1        201         4    yes     no   yes        yes   -1.088714   .0637559
    137   4.919981   4.994249          0        294         2    yes     no   yes        yes   -.1049232   .0636456
     94   4.543295   4.864575          0        263         3    yes     no   yes        yes   -.4538165   .0633344
     53   3.970292   4.227365          1        263         5    yes     no    no        yes   -.3622662   .0589003
    110    4.70048   4.162444          1        263         0    yes     no    no        yes    .757655    .057552
     32   3.465736   4.261548          1        201         1    yes    yes   yes         no   -1.117462   .0521669
     99    4.59512   4.592202          0        294         1    yes    yes    no         no    .0040933   .0505488
     23   3.135494   4.120585          1        185         0    yes    yes    no         no   -1.381421   .0496652
     40   3.688879   4.406658          1        263         0     no    yes   yes        yes   -1.004869   .0464619

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

. tab3way twin rp dyst


Table entries are cell frequencies
Missing categories ignored

-----------------------------------
          | 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
- 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 $n \geq 120$)
    - $n < 120$ only values beyond +/- 1

- wpc model

```
. summ cook dfit

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        cook |      1574    .0006351    .0013679   5.57e-10   .0174304
        dfit |      1574    .0003931    .0797594  -.4179343   .3664328


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

. count if cook>=.0025413 /* many (n=92) high Cook's D values */
   92

. 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
```

- cows with twins and/or reproductive disease were mildly influential, but no reason to delete them

- Delta-beta (or DFBETA)
  - influence of an observation on a specific regression coefficient (β)
  - for obs. "i" and predictor x_j
    - $DFBETA_{ij} = \frac{(\beta_j - \beta_{j(i)})}{SE}$
    - 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?
  - $DFBETA_{ij} > 0$ (<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

- wpc model

```
. sum dfb_hs2 dfb_dyst dfb_rp dfb_vd dfb_int

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     dfb_hs2 |      1574    9.10e-08    .0239995     -.1043   .1487626
    dfb_dyst |      1574   -.0000162    .0299013  -.2078438   .2472617
      dfb_rp |      1574    4.82e-07    .0254968  -.1650124   .1895015
      dfb_vd |      1574    2.40e-06     .029381  -.2821976   .3064255
-------------+--------------------------------------------------------
     dfb_int |      1574   -7.50e-06    .0236208  -.2360168   .2135575

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

. count if abs(dfb_dyst)>.05041127        /* n=74 */
   74
```

  - large values are cases of the correspondent predictor (eg. cows with dystocia)
  - linear effect of herd size much larger influence than quadratic term.
    - for small and large herds

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

```
-------------------------------------
    Variable |     ln        ln_noout
-------------+-----------------------
    aut_calv | -.13716214   -.13695147
       hsize |  .34365799    .33916323
      hsize2 |  .21187276    .20269649
     parity1 |  .01298436     .0113333
        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=0 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
  - influential diagnostics
    - similar to log-transformed model
    - more in VER.

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

Stata code

```
* VHM 812 - Winter 2014
* 2bl - Linear regression: diagnostics

* change working directory and open a log file
cd "F:\javier\Presentations\TA\data_epi_teach"

* capture log close
* log using 2bl_reg_dx.log, text replace
set more off

* open the DAISY dataset
use daisy2red.dta, clear
gen parity1=parity-1
gen calv_mth=month(calv_dt)
gen aut_calv=(calv_mth>=2 & calv_mth<=7) if !missing(calv_mth)
gen hsize=(herd_size-251)/100
gen hsize2=hsize^2

*final model
reg wpc hsize hsize2 parity1 aut_calv twin dyst i.rp##vag_disch

*initial diagnostic
capture drop fit stdres /* adjust if not defined */
predict fit, xb
predict stdres, rstandard

*homoskedasticity
scatter stdres  fit
estat hettest
estat imtest

*normality
qnorm stdres
hist stdres, normal
swilk  stdres


* transforming wpc
boxcox wpc aut_calv hsize hsize2 parity1 twin dyst rp vag_disch
gen wpc_ln=ln(wpc)
regress wpc_ln aut_calv hsize hsize2 parity1 twin dyst rp##vag_disch

*initial model checking
capture drop fit stdres /* adjust if not defined */
predict fit, xb
predict stdres, rstandard
*homoskedasticity
estat hettest
estat imtest
scatter stdres  fit
*normality
qnorm stdres
hist stdres, normal
swilk  stdres

* check for collinearity issues
estat vif
estat vce, corr
```

```
**linearity
* check modelling of herd_size and parity
lowess stdres herd_size
lowess stdres parity

* looking at herd_size without herd_size in the model
regress wpc_ln aut_calv parity twin dyst rp##vag_disch
predict stdres1, rstandard
lowess stdres1 herd_size

* looking at parity without parity in the model
regress wpc_ln aut_calv hsize hsize2 twin dyst rp##vag_disch
predict stdres2, rstandard
lowess stdres2 parity

* estimates and CIs backtransformed from ln-scale
regress wpc_ln aut_calv hsize hsize2 parity1 twin dyst rp##vag_disch
margins , over(rp vag_disch dyst) at(parity1=0 aut_calv=0 hsize=0 hsize2=0 twin=0)
expression(exp(predict(xb)))
marginsplot, bydimension(dyst)


* residuals and diagnostics
regress wpc_ln aut_calv hsize hsize2 parity1 twin dyst rp##vag_disch

capture drop fit stdres /* adjust if not defined */
predict fit, xb
predict stdres, rstandard
predict delres, rstudent
predict lev, lev
predict cook, cooksd
predict dfit, dfit
predict dfb_hs2, dfbeta(hsize2)
predict dfb_dyst, dfbeta(dyst)
predict dfb_rp, dfbeta(1.rp)
predict dfb_vd, dfbeta(1.vag_disch)
predict dfb_int, dfbeta(1.rp#1.vag_disch)
scalar nobs=1574
scalar nparam=10

* examine outliers
* number of standardized residuals outside -2/+2 and -3/+3
* expect 5% (n=79) outside -2,+2 and 1% (n=14) outside -3,+3
count if abs(stdres)>2 /* n=50 */
count if abs(stdres)>3 /* n=6, so fewer very extreme residual values than expected
*/
sum delres, d
* outlier test based on deletion residuals
display 2*nobs*ttail(nobs-nparam-1, 5.02) /* P=0.0009 so S */
display invttail(nobs-nparam-1,.025/nobs)
**there are two observations with extreme residual values (outliers)
** two cows with 1 days interval from the end of wp to conception
** cows 2272 (herd 106) and cow 1032 (herd 4)

* browse most extreme residuals
sort stdres
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres in 1/10 /* most extreme negative residuals */
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres in -10/-1 /* most extreme positive residuals */
```

```
sort delres
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres in 1/5 /* most extreme negative residuals */
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres in -5/-1 /* most extreme positive residuals */


* leverage and influence diagnostics
summarize lev cook dfit
display "leverage cutoff:  " 2*nparam/nobs
display "conservative leverage cutoff:  " 3*nparam/nobs
display "Cook's D cutoff:   " 4/nobs
display "DFITS cutoff:   " 2*sqrt(nparam/nobs)*(nobs>=120)+1*(nobs<120)

* large leverage values
count if lev>=.01905972 /* many (n=108) high leverage values */
gsort -lev
browse cow wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres lev in 1/10 /* most extreme leverages */
list wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres lev in
1/10, clean noobs
summ aut_calv herd_size parity1 twin dyst rp vag_disch

* large Cook's D values
summ cook dfit
count if cook>=.0025413 /* many (n=92) high Cook's D values */
gsort -cook
browse cow wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres cook dfit in 1/10
list wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres lev
cook in 1/10, clean noobs /* most extreme Cook's D values */

* large DFITS values
count if abs(dfit)>=.15941443 /* many (n=92) high DFITS values */
sort dfit
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres lev
cook dfit in 1/10 /* most extreme negative DFITS values */
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres lev
cook dfit in -10/-1 /* most extreme positive DFITS values */

* dfbeta diagnostics (for a subset of predictors)
sum dfb_hs2 dfb_dyst dfb_rp dfb_vd dfb_int //only for hsize2 since hsize will be
quite collinear and difficult to interpret//
display "DFBETA cutoff:   " 2/sqrt(nobs)*(nobs>=120)+1*(nobs<120)
count if abs(dfb_dyst)>.05041127    /* n=74 */
sort dfb_dyst
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres lev cook dfit dfb_dyst in 1/10 /* most extreme negative dfb_dyst values */
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres lev cook dfit dfb_dyst in -10/-1 /* most extreme positive dfb_dyst values */
* same approach for other DFBETAs

* dfbeta diagnostic for hsize2
sort dfb_hs2
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres lev cook dfit dfb_hs2 in 1/10 /* most extreme negative dfb_dyst values */
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres lev cook dfit dfb_hs2 in -10/-1 /* most extreme positive dfb_dyst values */

* dfbeta diagnostic for hsize
* ideally we would need to create two terms that are independent in order to assess
```

```
dfbeta for hsize and hsize2
* I am not going to do this here, instead I will fit a model with only hsize and
look at the dfbeta for the linear term
regress wpc_ln aut_calv hsize parity1 twin dyst rp##vag_disch
predict dfb_hs, dfbeta(hsize)
summ dfb_hs
sort dfb_hs
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres lev
cook dfit dfb_hs in 1/10 /* most extreme negative dfb_dyst values */
browse wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres lev
cook dfit dfb_hs in -10/-1 /* most extreme positive dfb_dyst values */

*analysis outliers observations
foreach var in delres lev dfit cook dfb_*{
format `var' %4.3f
}
l wpc wpc_ln  fit herd_size parity dyst rp vag_disch if wpc==1, clean compress
noobs
l wpc wpc_ln  delres lev dfit cook dfb_dyst dfb_rp dfb_vd dfb_int dfb_hs if wpc==1,
clean compress noobs

*re-fit without outliers/influential obs
regress wpc_ln aut_calv hsize hsize2 parity1 twin dyst rp##vag_disch
estimate store ln
estat hettest
estat imtest
capture drop fit_group
egen fit_group=cut(fit), group(10)
tabstat stdres, statistics( mean sd  count ) by(fit_group)
sort delres
browse cow wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres cook dfit in 1/10
sort dfit
browse cow wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres cook dfit in 1/10
gsort -cook
browse cow wpc wpc_ln fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
delres cook dfit in 1/10
*the two outliers don't have the highest infl. values

regress wpc_ln aut_calv hsize hsize2 parity1 twin dyst rp##vag_disch if wpc~=1
estimate store ln_noout
estat hettest
estat imtest

estimate table ln ln_noout //models are very similar since outliers did not have
much influence

*other possible transformation sqroot /*recommended in VER*/
gen wpc_sqrt=sqrt(wpc)
regress wpc_sqrt aut_calv hsize hsize2 parity1 twin dyst rp##vag_disch
estimate store sqrt
estat hettest
estat imtest
capture drop fit stdres
capture drop delres
capture drop lev
capture drop dfit
capture drop cook
predict fit, xb
predict stdres, rstandard
```

```
predict delres, rstudent
predict lev, lev
predict cook, cooksd
predict dfit, dfit

summ stdres delres
capture drop fit_group
egen fit_group=cut(fit), group(10)
tabstat stdres, statistics( mean sd count ) by(fit_group)
scatter stdres fit
hist stdres, normal

*obtain predictions
margins , over(rp vag_disch dyst) ///
      at(parity=2 aut_calv=0 hsize=0 hsize2=0 twin=0) expression(predict(xb)^2)

matrix define marg_sqrt = r(table)'
capture drop marg_s*
svmat marg_sqrt
drop marg_sqrt2 - marg_sqrt9

*se predictions
capture drop var se
capture drop loci* upci* estwpc*
matrix define var=vecdiag(e(V))'
svmat var
gen se=sqrt(var)
scalar tstar=invttail(1564,.025)
gen loci_sqrt=sqrt(marg_sqrt1)-tstar*se
gen upci_sqrt=sqrt(marg_sqrt1)+tstar*se
gen estwpc_l=loci_sqrt^2
gen estwpc_u=upci_sqrt^2

*influential analysis
summ lev dfit cook
count if lev>=.01905972 /* many (n=108) high leverage values */
gsort -lev
browse cow wpc wpc_sqrt fit aut_calv herd_size parity1 twin dyst rp vag_disch
stdres delres lev in 1/10 /* most extreme leverages */
* large Cook's D values
summ cook dfit
count if cook>=.0025413 /* many (n=90) high Cook's D values */
gsort -cook
browse cow wpc wpc_sqrt fit aut_calv herd_size parity1 twin dyst rp vag_disch
stdres delres cook dfit in 1/10
count if abs(dfit)>=.15941443 /* many (n=92) high DFITS values */
sort dfit
browse wpc wpc_sqrt fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
lev cook dfit in 1/10 /* most extreme negative DFITS values */
browse wpc wpc_sqrt fit aut_calv herd_size parity1 twin dyst rp vag_disch stdres
lev cook dfit in -10/-1 /* most extreme positive DFITS values */
* need to estimate dbetas and analyze results
* re-fit a model without influential observations
* for instance model with obs -.2> dfit <.2
count if abs(dfit)>.2
regress wpc_sqrt aut_calv hsize hsize2 parity1 twin dyst rp##vag_disch if dfit>-.2
& dfit <.2
estimate store sqrt_dfit02
estimate table sqrt sqrt_dfit02, star(0.05 0.01 0.001)
margins , over(rp vag_disch dyst) ///
      at(parity=2 aut_calv=0 hsize=0 hsize2=0 twin=0) expression(predict(xb)^2)
```

```
*predictions for ln model
regress wpc_ln aut_calv hsize hsize2 parity1 twin dyst rp##vag_disch
margins , over(rp vag_disch dyst) at(parity=2 aut_calv=0 hsize=0 hsize2=0 twin=0)
expression(exp(predict(xb)))
matrix define marg_ln = r(table)'
capture drop marg_ln*
svmat marg_ln
drop marg_ln2 - marg_ln9

*se predictions
capture drop var se
capture drop ln_loci ln_upci ln_estwpc*
matrix define var=vecdiag(e(V))'
svmat var
gen se=sqrt(var)
scalar tstar=invttail(1564,.025)
gen ln_loci_sqrt=ln(marg_ln1)-tstar*se
gen ln_upci_sqrt=ln(marg_ln1)+tstar*se
gen ln_estwpc_l=exp(ln_loci)
gen ln_estwpc_u=exp(ln_upci)

*format numeric values to 3 decimal places
foreach var in marg_sqrt1 estwpc_l estwpc_u marg_ln1 ln_estwpc_l ln_estwpc_u{
format `var' %5.3f
}
list marg_sqrt1 estwpc_l estwpc_u marg_ln1 ln_estwpc_l ln_estwpc_u in 1/8, clean
header noobs
```