

Lecture 2b: Linear Regression Diagnostics (VER 14.8-10)

Index	Page
Example: WPC model (daisy2red.dta).....	2
Evaluating major assumptions.....	3
Transformation of Y (L1a).....	4
Re-assessing assumptions.....	4
Evaluating individual observations.....	6
Leverage (h).....	7
Influence diagnostics: Cook's distance and DFITS.....	9
Delta-beta (or DFBETA).....	11
What to do with Outliers/Influential observations.....	13
Other transformations	14

- Learning objectives

- ★ Assess assumptions linear regression
- ★ Identify outliers and influential observations
- ★ Develop approach linear reg. diagnostics

- Exercises

- ★ work on linear regression exercise 3

- Quiz #2

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: i.rp##i.vag_disch

- Final (candidate) model

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

Source	SS	df	MS	Number of obs	=	1,574
Model	296062.694	9	32895.8549	F(9, 1564)	=	13.22
Residual	3892027.86	1,564	2488.50886	Prob > F	=	0.0000
				R-squared	=	0.0707
				Adj R-squared	=	0.0653
Total	4188090.56	1,573	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					
yes	-8.263839	2.537751	-3.26	0.001	-13.24159 -3.286086
twin					
yes	20.68314	9.845165	2.10	0.036	1.37203 39.99425
dyst					
yes	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

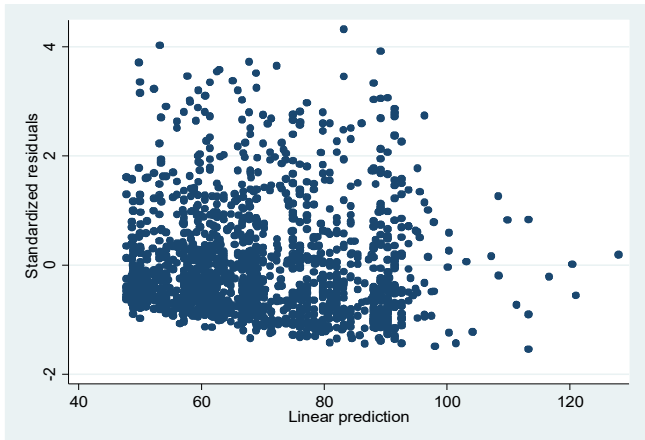
```
. hettest

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

chi2(1)      =    20.58
Prob > chi2   =    0.0000

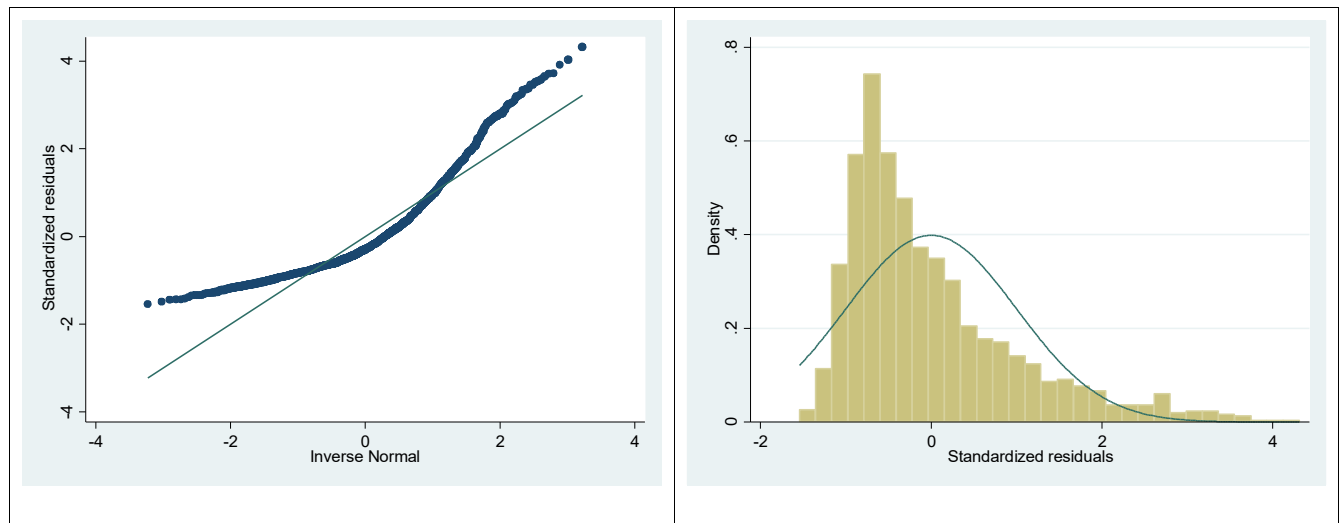
. imtest

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 (L1a)

- Box-cox
- Interpretation

Re-assessing assumptions

- Log-transformed model

```
xi:boxcox wpc hs100_ctr hs100_ctr_sq parity1 i.aut_calv i.twin i.dyst  
i.rp*i.vag_disch
```

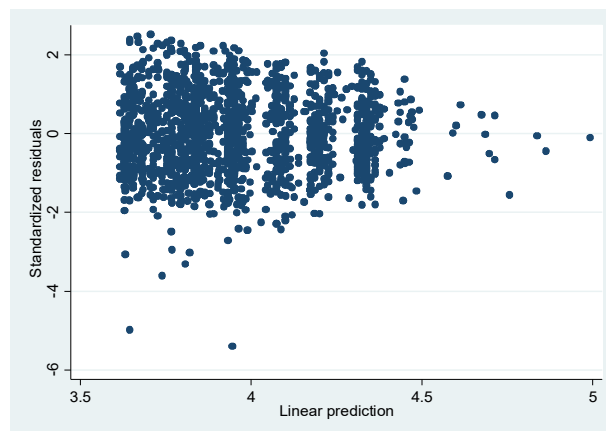
...output omitted...

```
Log likelihood = -7960.5368      Number of obs   =      1574  
                                LR chi2(8)         =      148.83  
                                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 \rightarrow log transformation

- Homoscedasticity (L1a: p11)

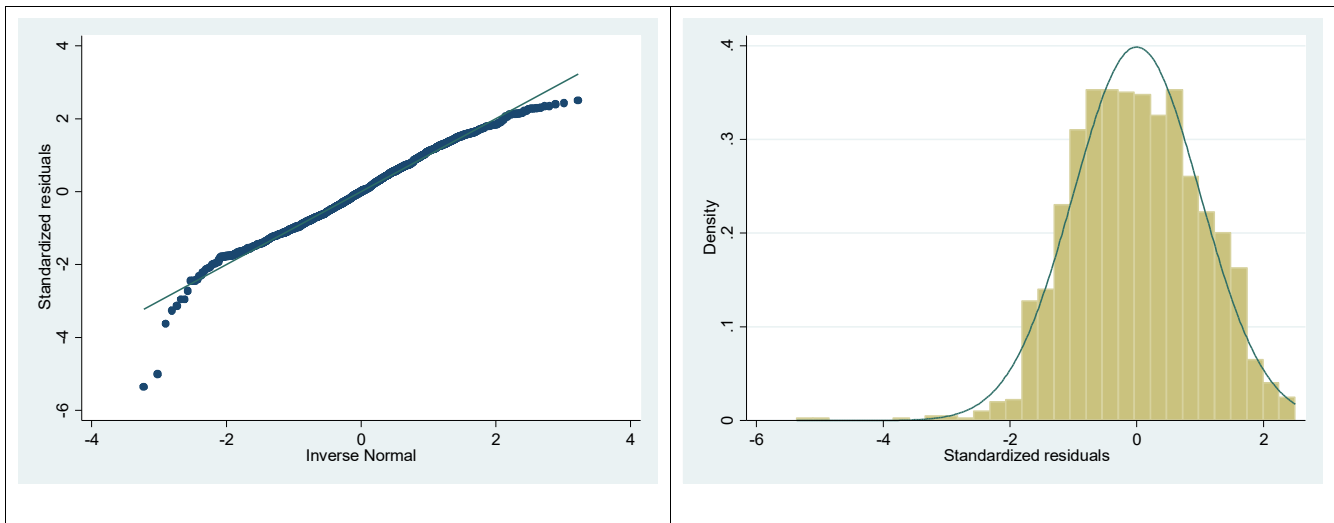


★ tests provide contradicting results

➔ imtest - constant variance

➔ however BPCW test highly significant

● Normality (L1a)



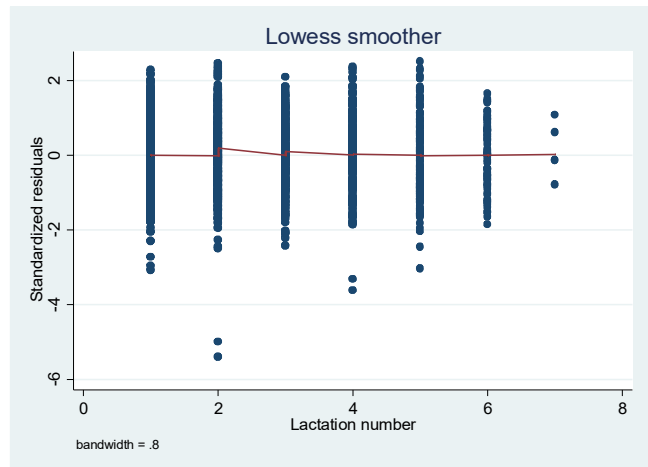
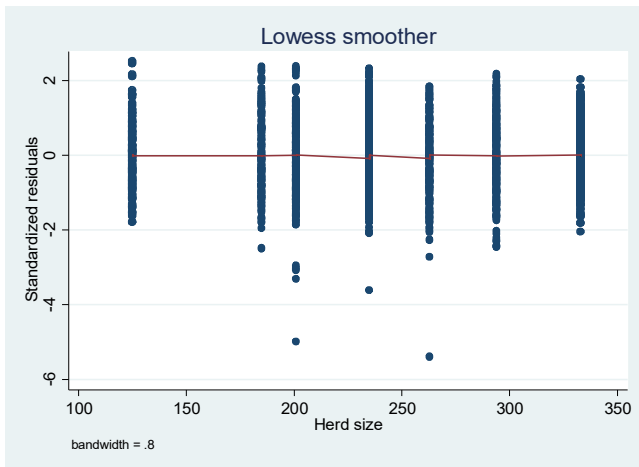
● Linearity (L1a)

★ 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)

- Large residuals
- Standardized residuals
 - ★ expect 5% > 2 and < -2
 - ★ expect 0.3% > 3 and < -3
- Deletion residuals
 - ★ t-test - $\text{prob} = 2 * n * t(\text{DFE}-1, d_i)$
- WPC log-transformed model - residuals
 - ★ standardized residuals
 - expected 5% = 79 (observed = 50)
 - expected 0.3% = 5 (observed = 6)
 - expected and observed values very similar
 - ★ deletion residuals
 - outlier cutoff value = 4.17
 - two observations \geq this cutoff point
 - indication that these are outliers
 - cows with $\text{wpc} = 1$
 - delete and refit the model (later)
 - only for diagnostics purposes

Leverage (h)

- 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 parameters (w/o intercept), n =# obs.
 - should also look for unusual observations independently of these cutpoints
- Stata - predict varname, leverage
 - ★ eg. predict lev_lnwpc, lev
 - ★ lists/graphs to assess values

● Computed after full model (ln_wpc)

. 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 */

. sort lev

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

wpc	wpc_ln	fit	aut_calv	herd_s~e	parity1	twin	dyst	rp	vag_di~h	stdres	lev
40	3.689	4.407	yes	263	0	no	yes	yes	yes	-1.005	0.046
23	3.135	4.121	yes	185	0	yes	yes	no	no	-1.381	0.050
99	4.595	4.592	no	294	1	yes	yes	no	no	0.004	0.051
32	3.466	4.262	yes	201	1	yes	yes	yes	no	-1.117	0.052
110	4.700	4.162	yes	263	0	yes	no	no	yes	0.758	0.058
53	3.970	4.227	yes	263	5	yes	no	no	yes	-0.362	0.059
94	4.543	4.865	no	263	3	yes	no	yes	yes	-0.454	0.063
137	4.920	4.994	no	294	2	yes	no	yes	yes	-0.105	0.064
45	3.807	4.577	yes	201	4	yes	no	yes	yes	-1.089	0.064
76	4.331	4.699	no	185	4	yes	no	yes	yes	-0.520	0.064

. table twin (rp dyst), nottotal

	Retained placenta at calving			
	no		yes	
	Dystocia at calving no	Dystocia at calving yes	Dystocia at calving no	Dystocia at calving yes
Twins born				
no	1,332	76	124	15
yes	15	2	9	1

★ large leverage values for less common combination of categorical predictors

Influence diagnostics: Cook's distance and DFITS

- Cook's distance and DFITS

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

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

★ 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 $\pm 2 * \sqrt{\frac{k+1}{n}}$ (if $n \geq 120$)

• $n < 120$ only values beyond ± 1

- Stata: `predict varname, cooksdfits`

★ `lists/graphs` to assess values

● Model: ln_wpc

★ 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=95) high Cook's D values */  
95
```

★ 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=95) high DFITS values */  
95
```

★ largest Cook's values

➔ not clear pattern - cows with at least one disease

★ largest negative DFIT values

➔ cows with twins and at least one of the diseases were the most influential

★ largest positive DFIT values

➔ cows without twins but with at least one of the diseases were the most influential

Delta-beta (or DFBETA)

★ influence of an observation on a specific regression coefficient (β)

★ for obs. "i" and predictor x_j

$$\rightarrow \text{DFBETA}_{j(i)} = \frac{(\beta_j - \beta_{j(i)})}{SE_j}$$

→ where $\beta_{j(i)}$ is the estimate of β_j without obs. "i"

★ measures change in β_j en terms of SE's

→ how many SE's does β_j increase(decrease) if "i" is dropped?

● cut-off values

★ $\text{DFBETA}_{j(i)} > 0$ (< 0) ~ obs. "i" pulls β_j up (down)

→ extreme values of $\text{DFBETA}_{j(i)}$ are notable

→ most meaningful for continuous predictors

★ concern if outside $\pm \frac{2}{\sqrt{n}}$ (for $n \geq 120$)

→ $n < 120$ only values beyond +/- 1

● Stata

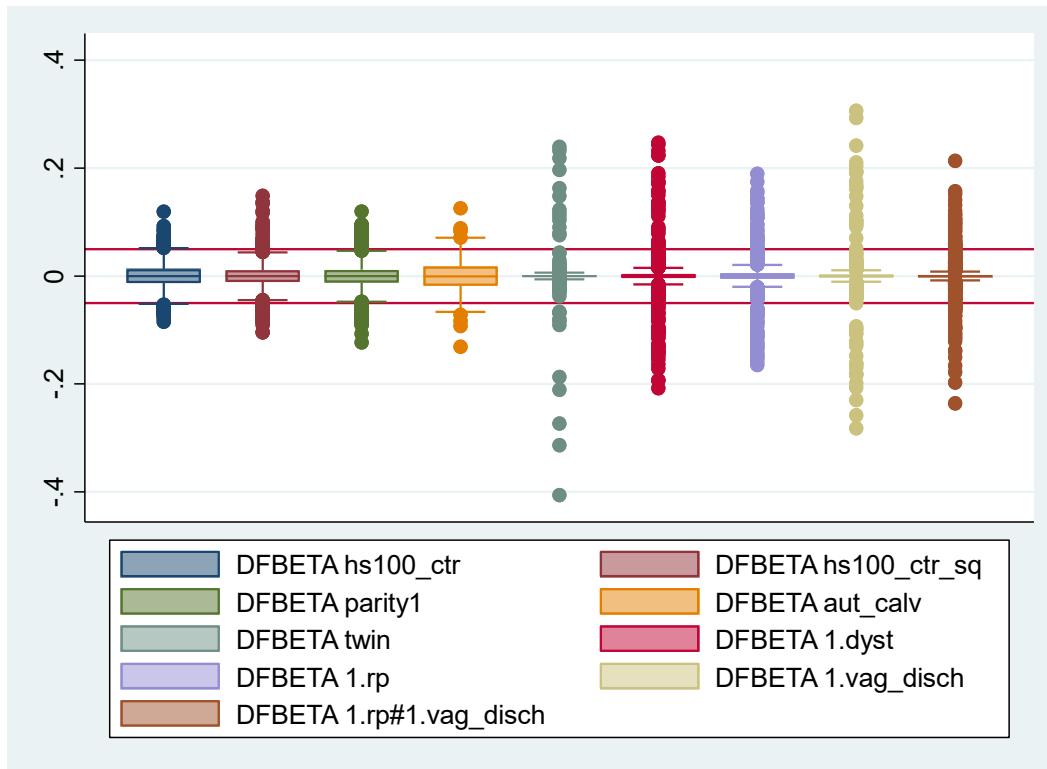
★ `predict varname, dfbeta(var in model)`

→ eg. `predict dfbdyst, dfbeta(dyst)`

★ `lists/graphs` to assess values

● Model: ln_wpc

```
. 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 diseases and twins)

★ 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(s) (observation(s)) are not errors (eg. data entry errors)
- May be very informative
 - ★ always examine observations and values of the predictors
- Determine if outliers are significant (L1a)
- Fit models with and without the outlier/influential observation(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 create an unrealistically good model

Other transformations

- See “l2b_linear_reg_diag_other_transf.do”
- 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~ln	fit	herd_size	parity	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~ln	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

rp#vag_disch#dyst	Original scale			Sqrt transformation			log-transformation		
	est	lo_ci	up_ci	est	lo_ci	up_ci	est	lo_ci	up_ci
no#no#no	66.605	62.220	70.990	58.267	54.583	62.071	50.127	47.005	53.456
no#no#yes	78.305	67.338	89.272	66.840	57.187	77.244	56.008	47.688	65.780
no#yes#no	67.833	53.767	81.899	58.064	46.695	70.671	48.840	39.737	60.027
no#yes#yes	79.533	63.311	95.755	66.622	52.636	82.255	54.570	43.017	69.225
yes#no#no	72.592	63.048	82.135	64.364	56.085	73.213	56.085	48.761	64.510
yes#no#yes	84.292	70.918	97.666	73.359	61.107	86.731	62.666	51.506	76.243
yes#yes#no	96.672	78.396	114.947	90.259	71.883	110.723	82.645	63.217	108.045
yes#yes#yes	108.372	87.603	129.141	100.857	78.877	125.534	92.342	68.098	125.216

*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.