

Regression Model Checking

February 11, 2020

Biostatistics 208

Model checking for linear regression:

- Linearity
- Normality
- Constant variance
- Influential points
- Covariate overlap

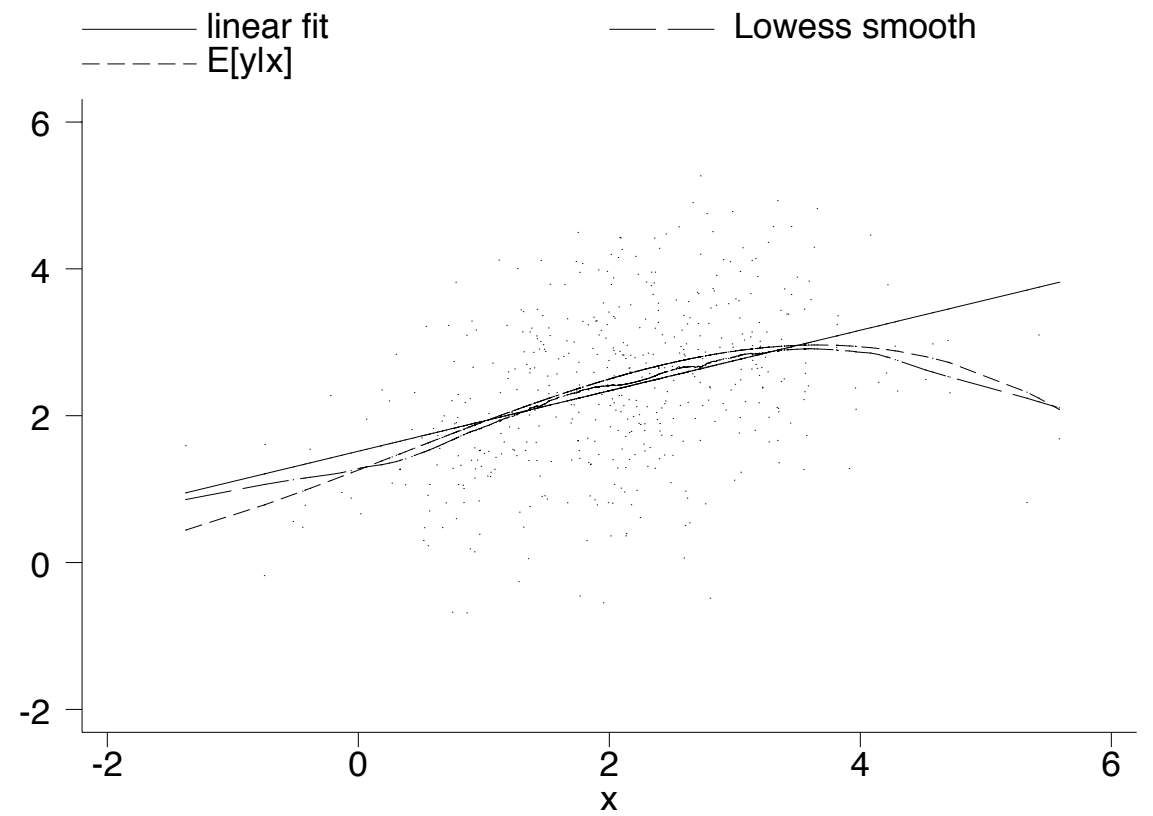
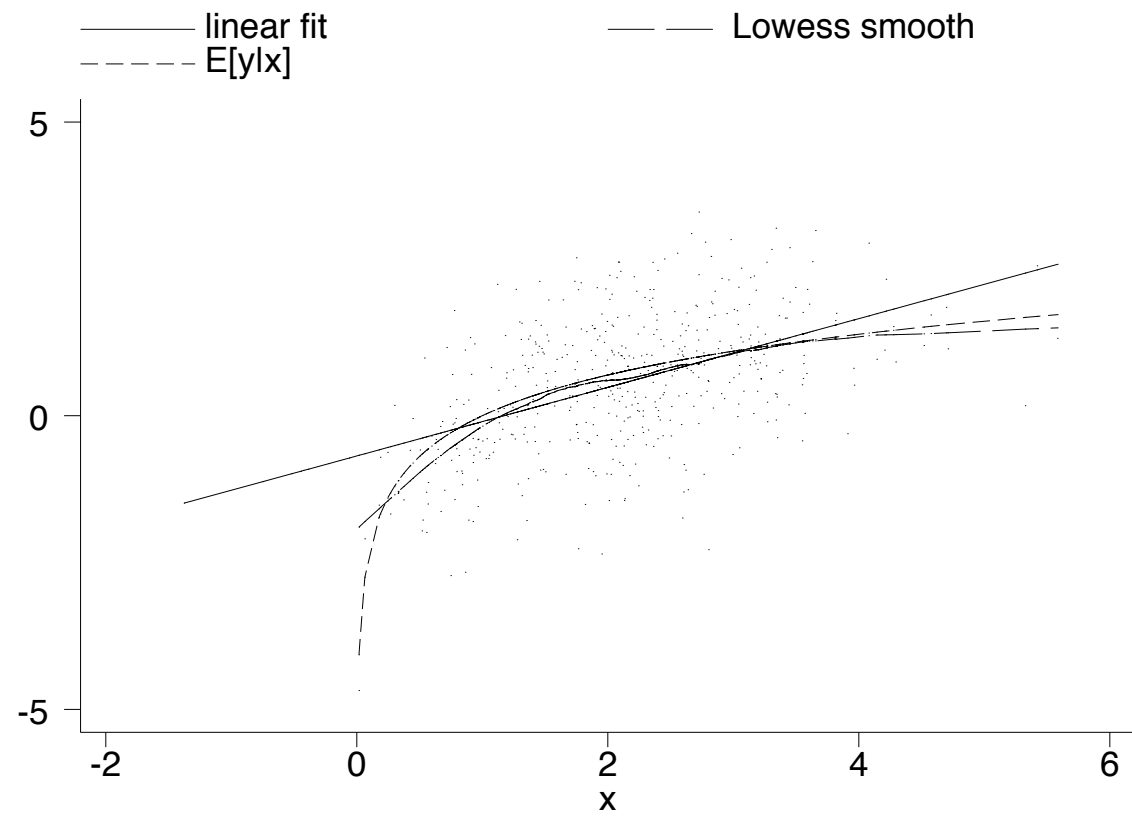
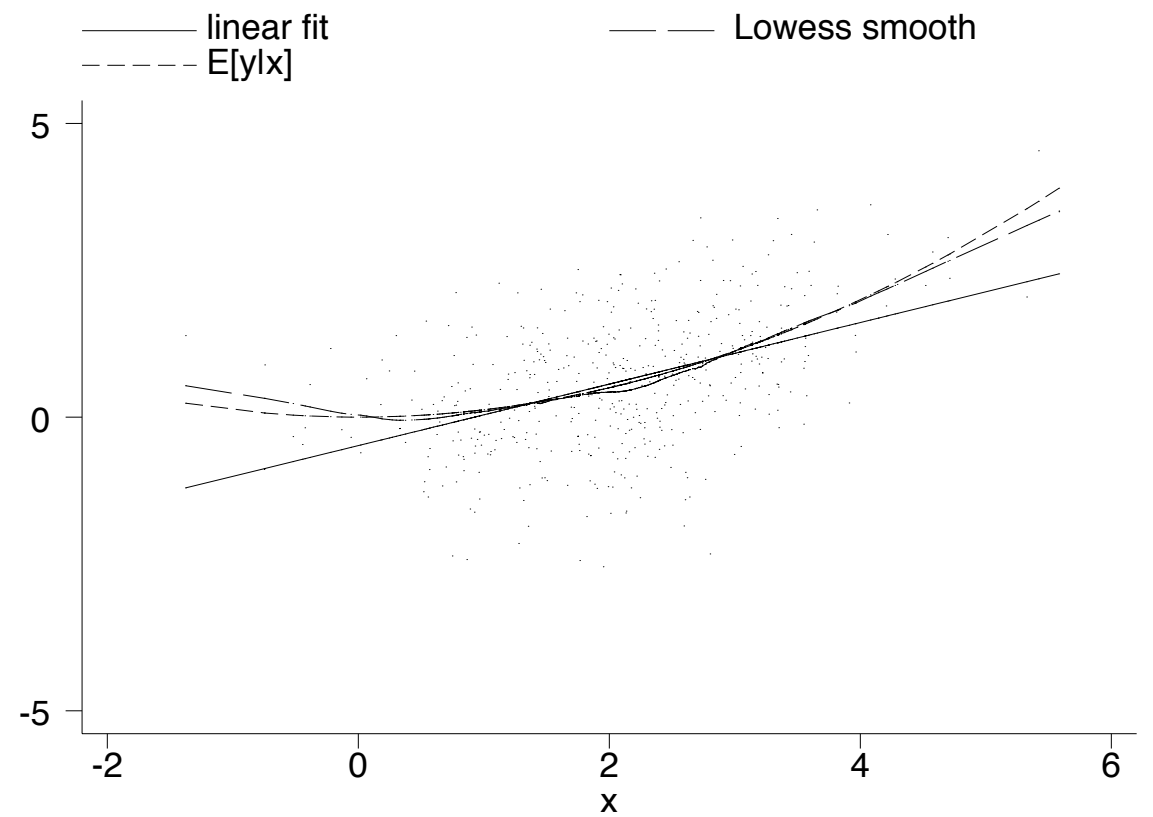
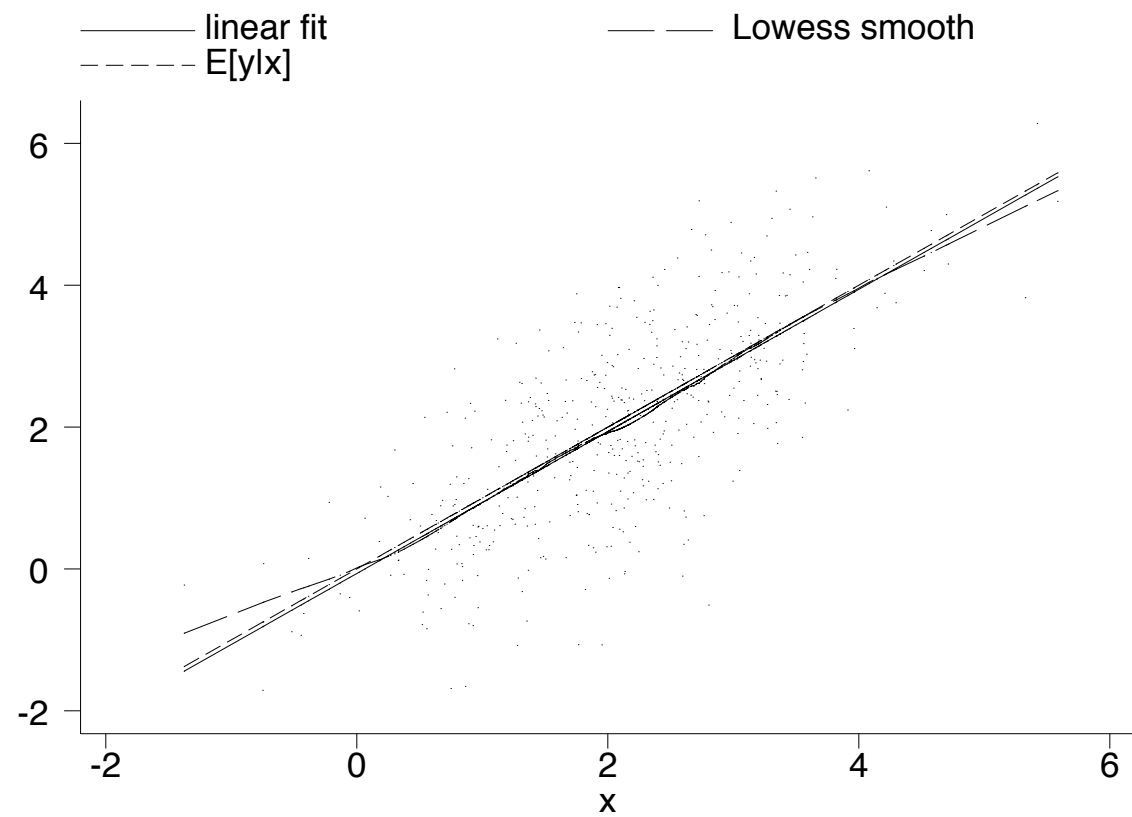
Model checking: linearity

- Regression relationship is assumed linear in continuous predictors
 - i.e., “slope” coefficients are constants
 - equivalently, assumes no curvature
- If model is correct
 - residuals have mean zero at every value of predictor
 - useful for diagnosing departures from linearity

Model checking: linearity

- If assumption badly violated, result can be
 - biased coefficient estimates, residual confounding
 - reduced precision and power, missed real effects
 - misleading, over-simplified conclusions

Three departures from linearity



Diagnostics for nonlinearity: RVP and CPR plots

- To account for effects of other predictors, diagnostics use residuals rather than outcome
- Basic approach: check for non-linear patterns in plots of residuals versus each continuous predictor (RVP) plots
- Better alternative: *component plus residual* (CPR) plots
 - component due to predictor added back into residual
 - results in *partial residual* for that variable
 - e.g. for a model with 2 predictors x_1 & x_2 , partial resid. for x_1 :

$$\underbrace{y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2)}_{\text{residual}} + \underbrace{\beta_1 x_1}_{\text{linear component}} = \underbrace{y - (\beta_0 + \beta_2 x_2)}_{\text{partial residual for } x_1}$$

- plot of partial residual vs x_1 reveals possibly un-modeled nonlinearity

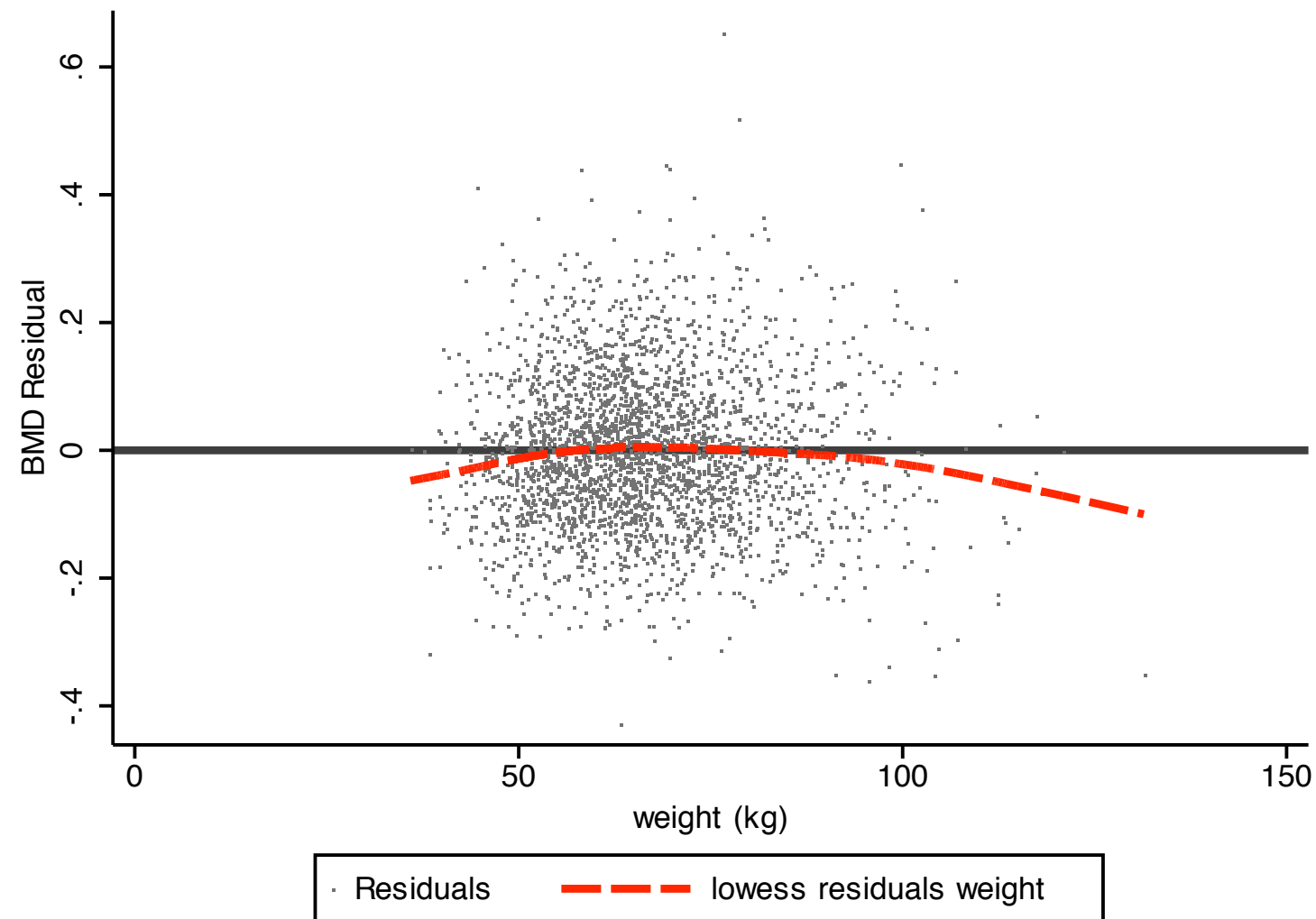
Diagnostics for nonlinearity: RVP and CPR plots

- CPR plots better for diagnosing non-linearity:
 - show trend, RVP plots do not
 - slightly easier to add LOWESS smooth
- Note: use RVP for quadratic, other polynomial models where predictor is included in multiple terms - e.g.:

$$E[y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

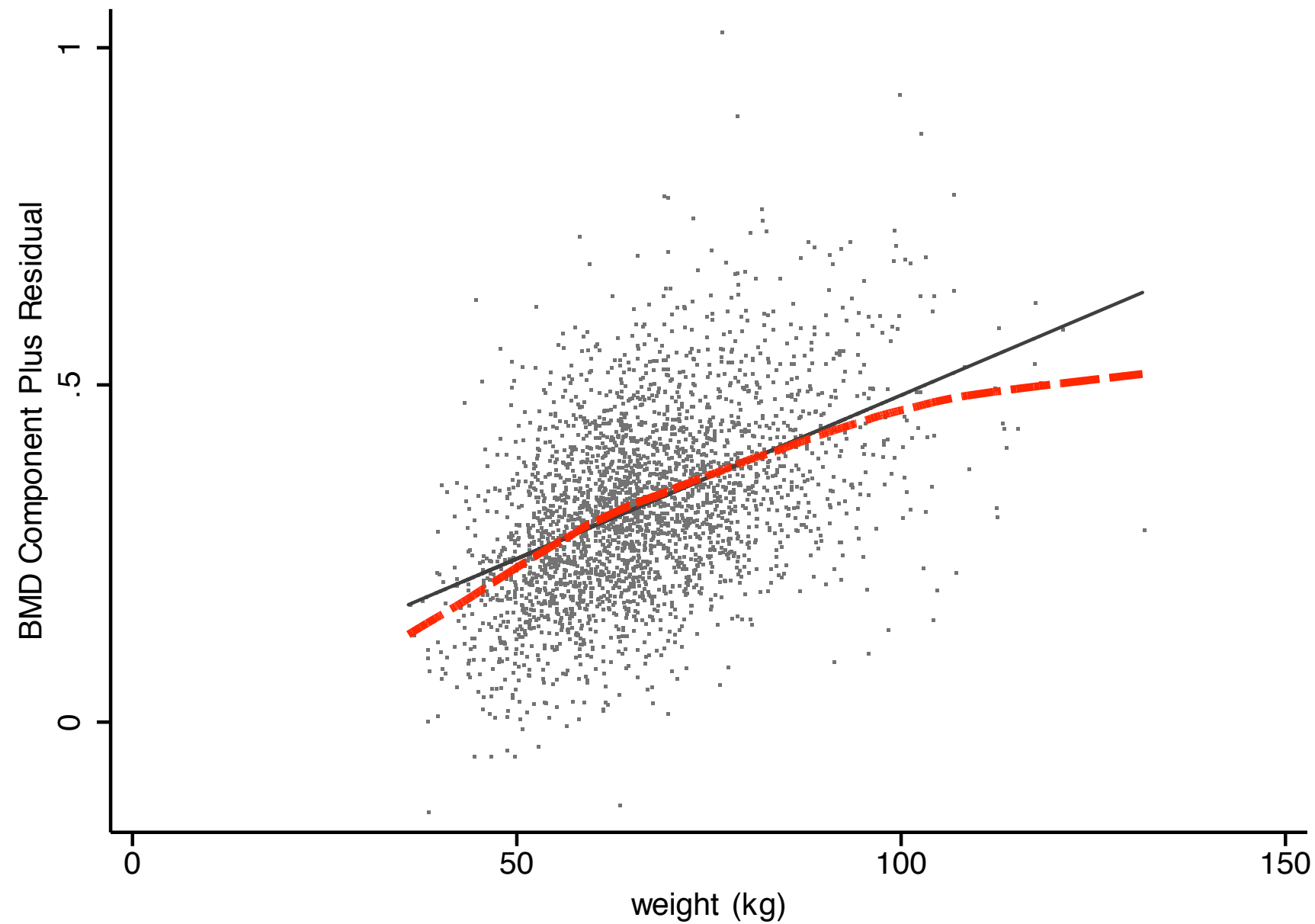
- In both CPR and RVP: mismatch of linear regression line, LOWESS smooth indicates lack of linearity

RVP plot for weight and BMD (SOF data)



Residuals from linear model for dependence of mean BMD on age & weight

CPR plot for weight and BMD



Partial residual for weight = overall residual + estimated linear component from weight

Stata code for RVP and CPR plots

```
regress bmd age weight  
predict residuals, resid
```

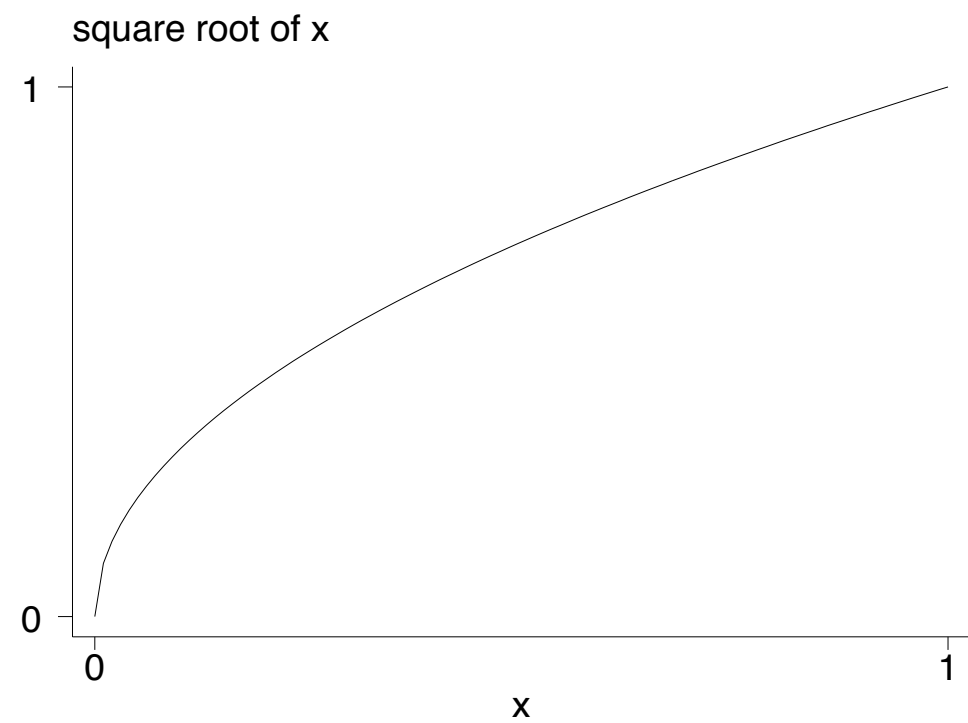
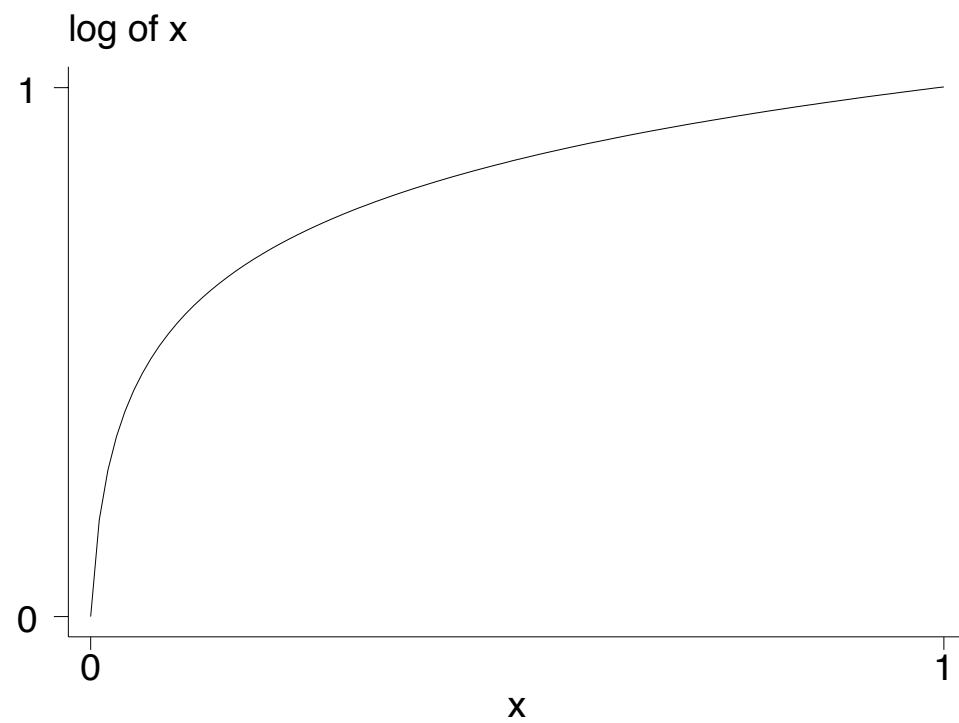
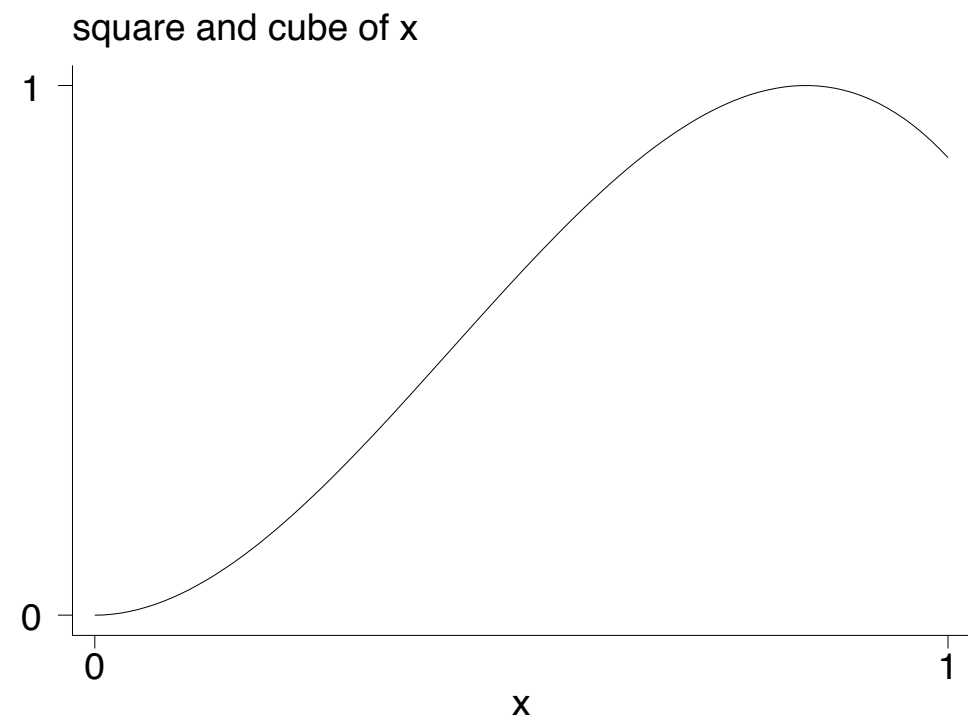
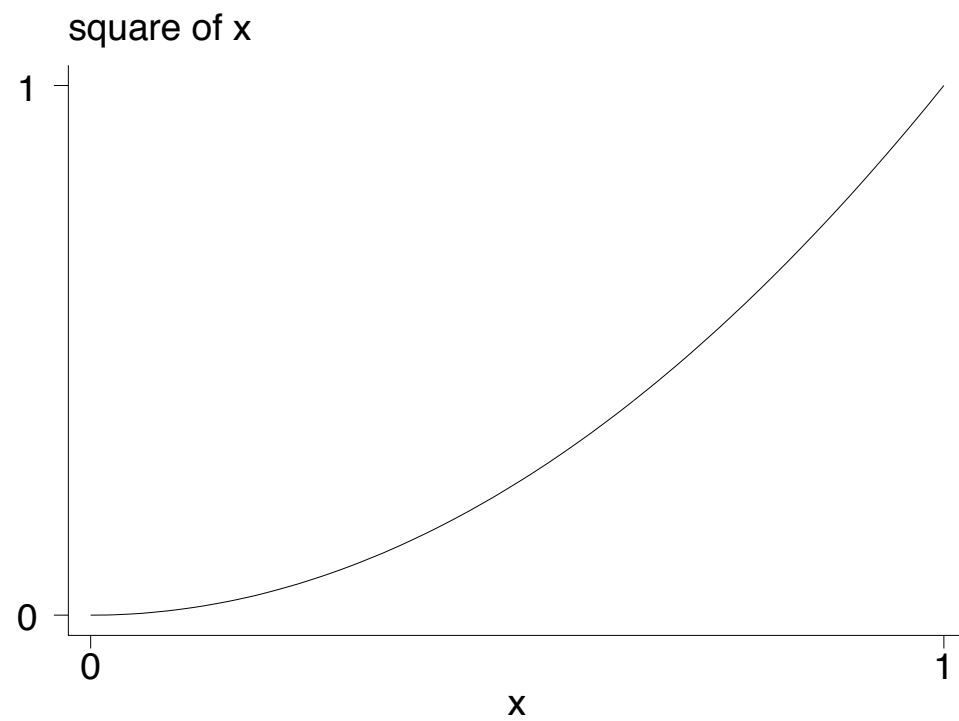
```
rvpplot weight, ///  
    lw(thick) msize(tiny) yline(0, lw(thick)) ///  
    addplot(lowess residuals weight, lw(thick) lcol(red) clpat(longdash)) ///  
    plotregion(style(none)) scheme(slmono) ytitle("BMD Residual") ///  
    xtitle("weight (kg)") xlabel(0(50)150)
```

```
cprplot weight, ///  
    rlopts(clpat(solid) lw(thick)) ///  
    lsopts(bw(.5) clpat(longdash) lw(thick) lcol(red)) ///  
    plotregion(style(none)) scheme(slmono) msize(tiny) ///  
    ytitle("BMD Component Plus Residual") xtitle("weight (kg)") ///  
    xlabel(0(50)150)
```

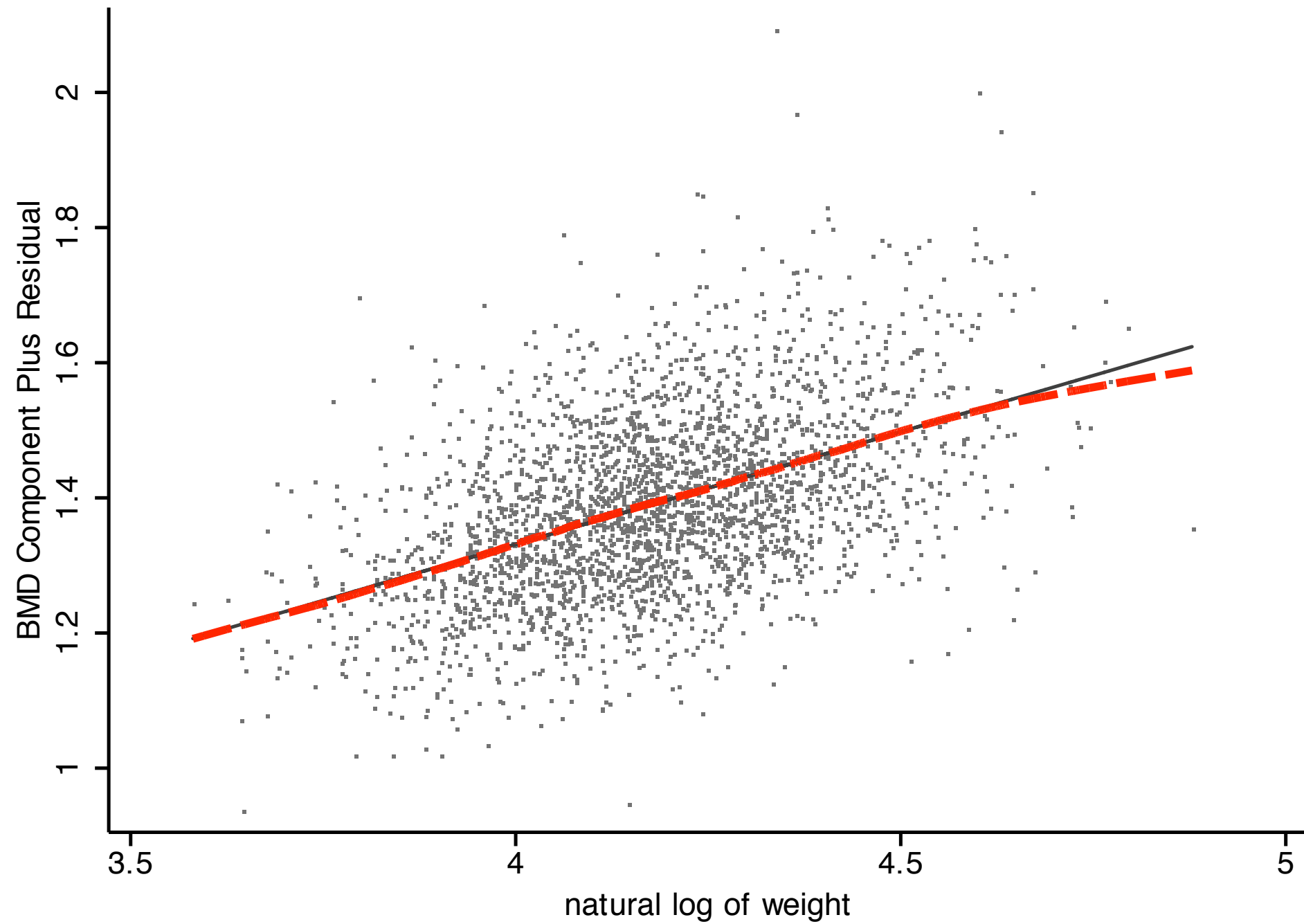
Solution: transform continuous predictors

- Smooth predictor transformations to fix non-linearity:
 - $\log(x)$ – provided $E[y | x]$ is “monotone”
 - square root, cube root, other fractional powers of x
 - polynomials: x^2, x^3 (lower order terms usually included in the model)

Predictor transformations



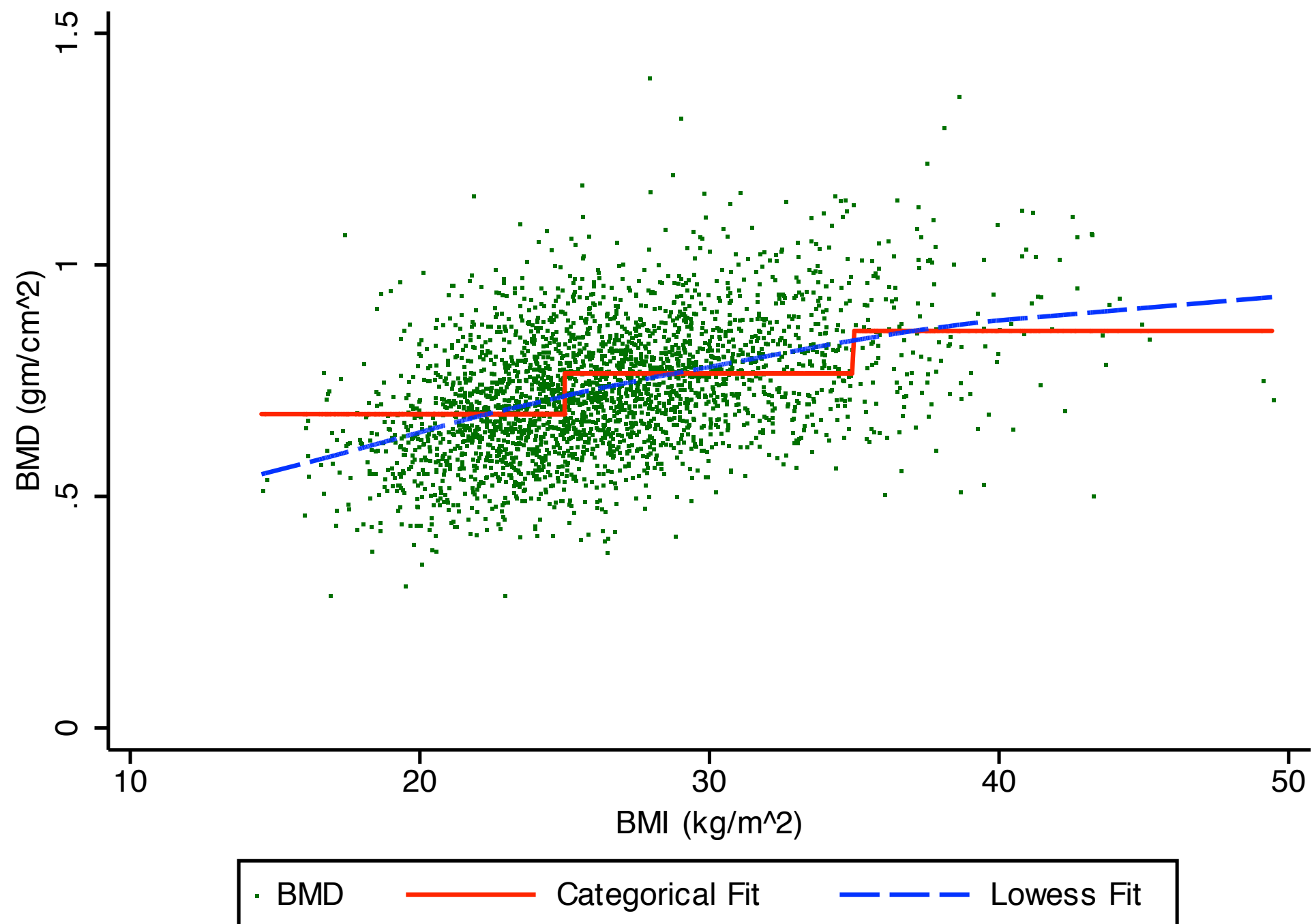
CPR plot for log-weight and BMD (SOF data)



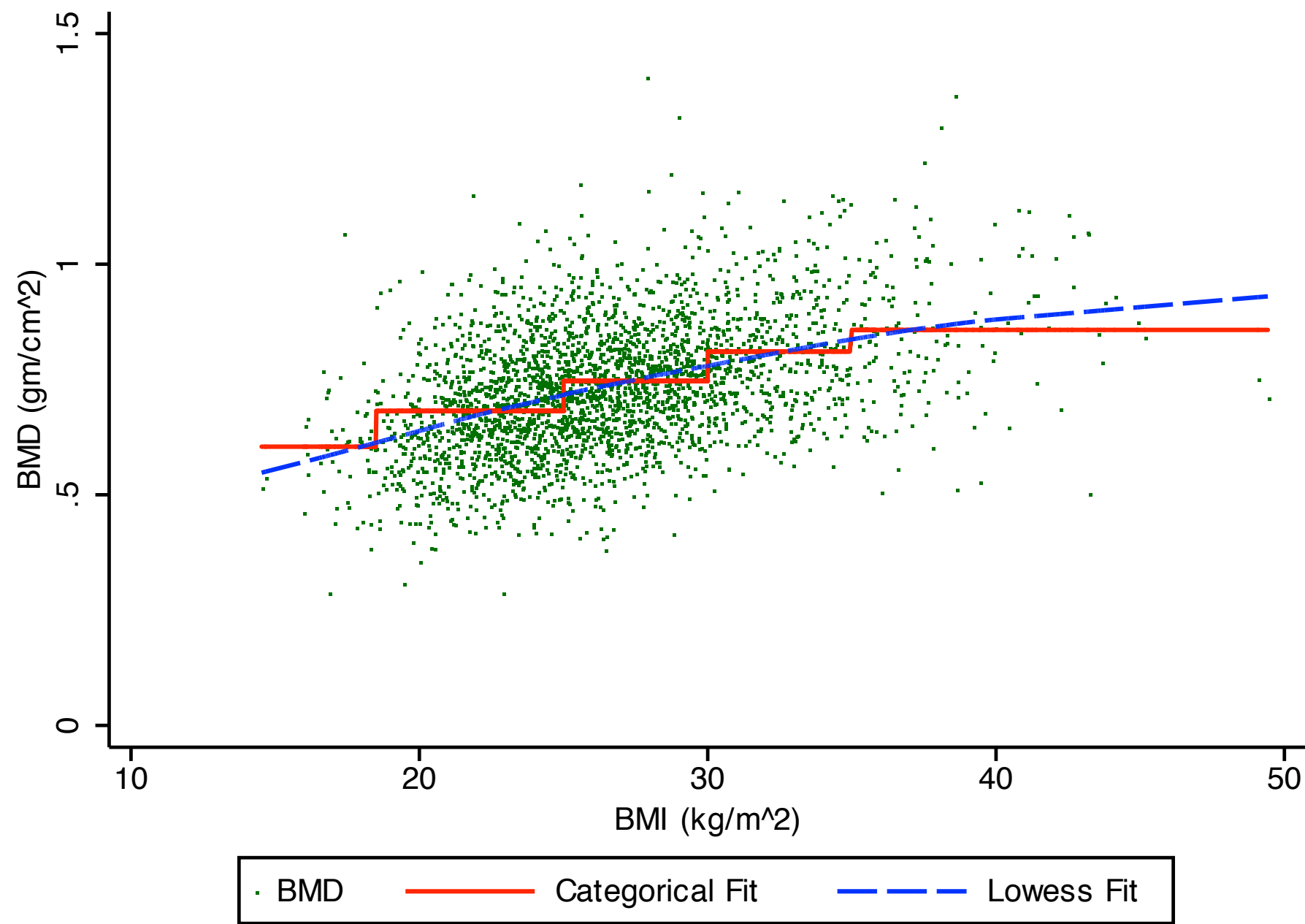
Alternatives: categorize the predictor

- Split at quantiles or clinically familiar cut-points
- Models mean as a “step function”
- Flexible, familiar, clinically interpretable, but
 - unrealistic if the regression line changes smoothly, sensitive to choice of cut-points, inefficient compared to smooth transformations
- Numbers of categories must balance fit against noisiness

Too coarsely categorizing the predictor



A better trade-off



Stata code for categorical transformations

```
recode bmi min/25=2 25/35=3 35/max=4, gen(bmicat)
regress bmd i.bmicat
predict fitted, xb
twoway ///
    (scatter bmd bmi, msize(vtiny)) ///
    (line fitted bmi, sort lpattern(solid) lcolor(red) lwidth(medthick)) ///
    (lowess bmd bmi, lpattern(longdash) lcolor(blue) lwidth(medthick)), ///
    plotregion(style(none)) scheme(s1color) ///
    ytitle("BMD (gm/cm2)") ///
    xtitle("BMI (kg/m2)") ///
    legend(label(2 "Categorical Fit") label(3 "Lowess Fit") rows(1))
recode bmi min/18.5=1 18.5/25=2 25/30=3 30/35=4 35/max=5, gen(bmicat2)
regress bmd i.bmicat2
predict fitted2, xb
twoway ///
    (scatter bmd bmi, msize(vtiny)) ///
    (line fitted2 bmi, sort lpattern(solid) lcolor(red) lwidth(medthick)) ///
    (lowess bmd bmi, lpattern(longdash) lcolor(blue) lwidth(medthick)), ///
    plotregion(style(none)) scheme(s1color) ///
    ytitle("BMD (gm/cm2)") ///
    xtitle("BMI (kg/m2)") ///
    legend(label(2 "Categorical Fit") label(3 "Lowess Fit") rows(1))
```

Alternatives: linear, restricted cubic splines

- *Linear splines*: piecewise linear with changes in slope at “knots”
 - coefficients interpretable as slope between knots
- *Restricted cubic splines*: smooth curve, cubic between knots
 - tails better behaved than polynomials - *restricted* to linearity before first and after last knot
 - easy tests for overall effects and non-linearity, but coefficients uninterpretable; presentation requires plotting
- Also: fractional polynomials (`fracpoly` command)

Ref. on cubic splines and fractional polynomials: <http://www.stata-journal.com/sjpdf.html?articlenum=st0120>

Linear spline model for BMI effect on BMD in SOF

* linear spline with knots at BMI = 18.5, 25, 30 & 35

```
. mkspline bmi1 18.5 bmi2 25 bmi3 30 bmi4 35 bmi5 = bmi
```

```
. regress bmd bmi1-bmi5
```

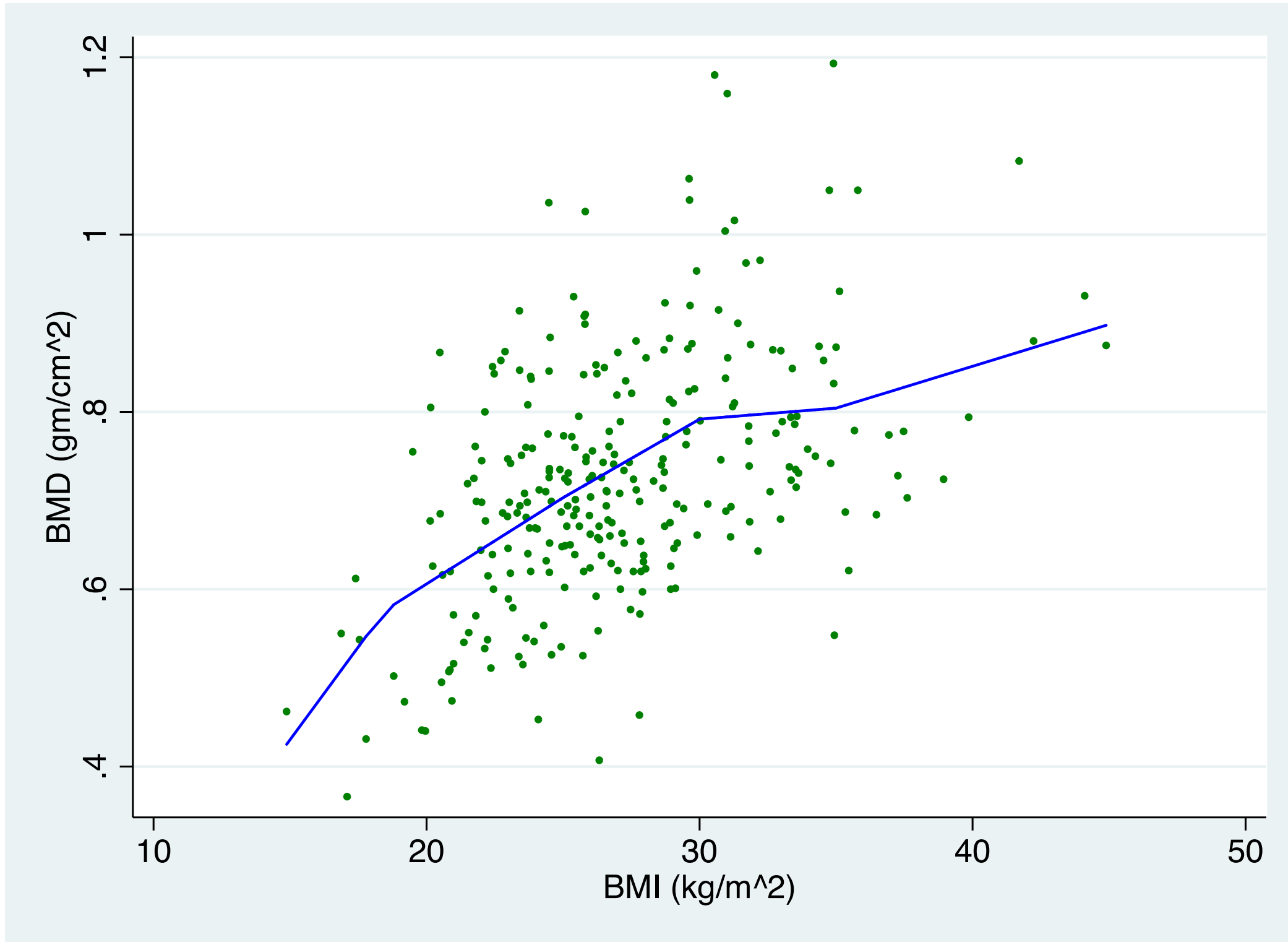
Source	SS	df	MS	Number of obs =	278
Model	1.34269169	5	.268538337	F(5, 272) =	18.91
Residual	3.86165215	272	.014197251	Prob > F =	0.0000
				R-squared =	0.2580
				Adj R-squared =	0.2444
Total	5.20434383	277	.018788245	Root MSE =	.11915

bmd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bmi1	.0418738	.0300524	1.39	0.165	-.017291	.1010387
bmi2	.0194547	.0060541	3.21	0.001	.0075358	.0313736
bmi3	.017719	.0054267	3.27	0.001	.0070354	.0284027
bmi4	.0024954	.0070065	0.36	0.722	-.0112986	.0162893
bmi5	.0094409	.007597	1.24	0.215	-.0055154	.0243972
_cons	-.1979034	.5417402	-0.37	0.715	-1.26444	.8686334

Coefficients for BMI give slopes for
lines segments between knots

Interpretation of the overall F statistic for this model?

Linear spline fit



Testing for non-linearity using linear splines

```
. testparm bmi*, equal
```

```
(1) -bmi1+bmi2=0
```

```
(2) -bmi1+bmi3=0
```

```
(3) -bmi1+bmi4=0
```

```
(4) -bmi1+bmi5=0
```

```
F(4, 272)= 2.24
```

```
Prob > F = 0.0654
```

- F statistic evaluates null hypoth. that specifies equality of slopes between knots
- rejection implies significant evidence for presence of non-linearity

Restricted cubic spline model for BMI effect on BMD in SOF

```
. mkspline bmispl = bmi, cubic
.* selects default knots and creates new predictors bmispl - bmispl4
. regress bmd bmispl*
```

Source	SS	df	MS	Number of obs	=	278
				F(4, 273)	=	25.21
Model	1.40393728	4	.350984321	Prob > F	=	0.0000
Residual	3.80040655	273	.013920903	R-squared	=	0.2698
				Adj R-squared	=	0.2591
Total	5.20434383	277	.018788245	Root MSE	=	.11799

bmd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bmispl1	.0367141	.0078805	4.66	0.000	.0211999	.0522283
bmispl2	-.1583327	.0606301	-2.61	0.010	-.2776947	-.0389707
bmispl3	.6911728	.2756232	2.51	0.013	.1485558	1.23379
bmispl4	-.8292426	.3414449	-2.43	0.016	-1.501442	-.1570429
_cons	-.1408691	.1695457	-0.83	0.407	-.4746523	.1929141

} spline terms

Note: first term in a restricted cubic spline fit represents the linear component (i.e. bmispl1 in example)

- Test for any BMI effect on BMD

```
. testparm bmisp*
```

```
( 1)  bmisp1 = 0
( 2)  bmisp2 = 0
( 3)  bmisp3 = 0
( 4)  bmisp4 = 0
```

```
      F(   4,   273) =    25.21
      Prob > F =    0.0000
```

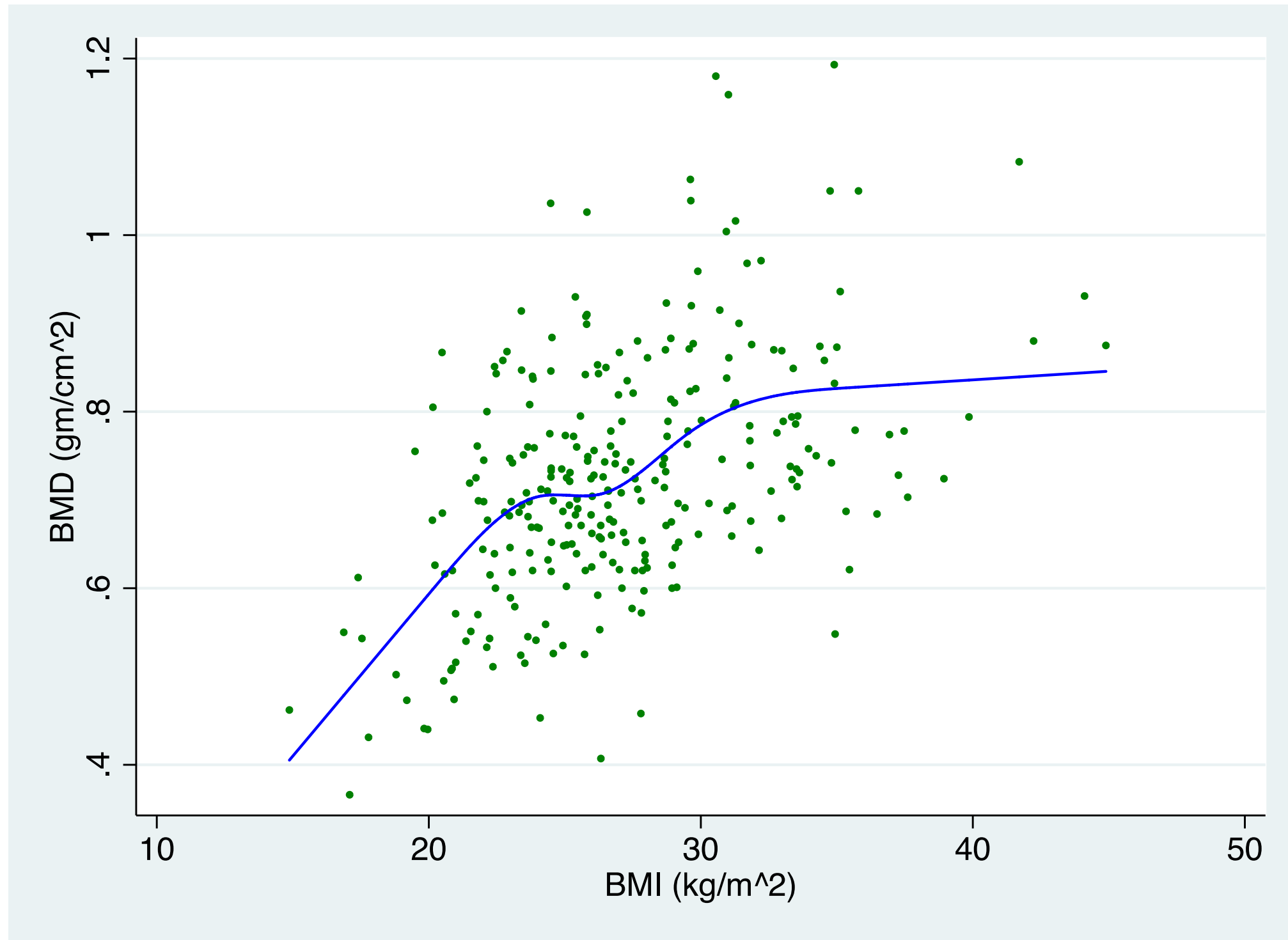
- Test for departure from linearity (involves all but the linear term)

```
. testparm bmisp2 bmisp3 bmisp4
```

```
( 1)  bmisp2 = 0
( 2)  bmisp3 = 0
( 3)  bmisp4 = 0
```

```
      F(   3,   273) =    4.51
      Prob > F =    0.0042
```


Cubic spline fit



Checking linearity: summary

- Diagnostics:
 - linear models: curved LOWESS smooth in CPR or RVP plot
 - more generally (i.e., linear, logistic, Cox models): fit restricted cubic spline, test for departure from linearity using `testparm` for all but first spline component
- Solutions: transform predictor, use linear or cubic splines
 - linear splines: piecewise linear; useful when linearity of particular segments is of interest
 - cubic splines: smooth, nonlinear functions

Model checking: normality

- Validity of t - and F -tests, CIs depend on assumption of normality of residuals (ε)
- Fairly robust to violations, especially short-tailed errors in larger samples
- However, long-tailed distributions of ε can degrade power, precision
- Diagnostics: Q-Q and other plots of residuals
 - tests for normality lack power where you need it

Example: model for mean LDL in HERS

```
. regress LDL BMI age nonwhite smoking drinkany
```

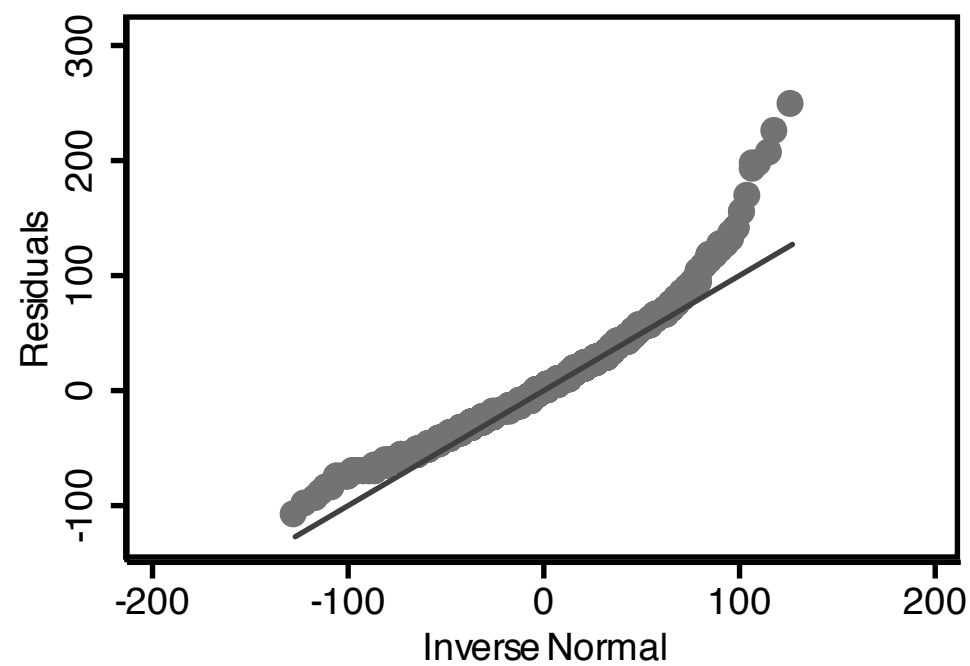
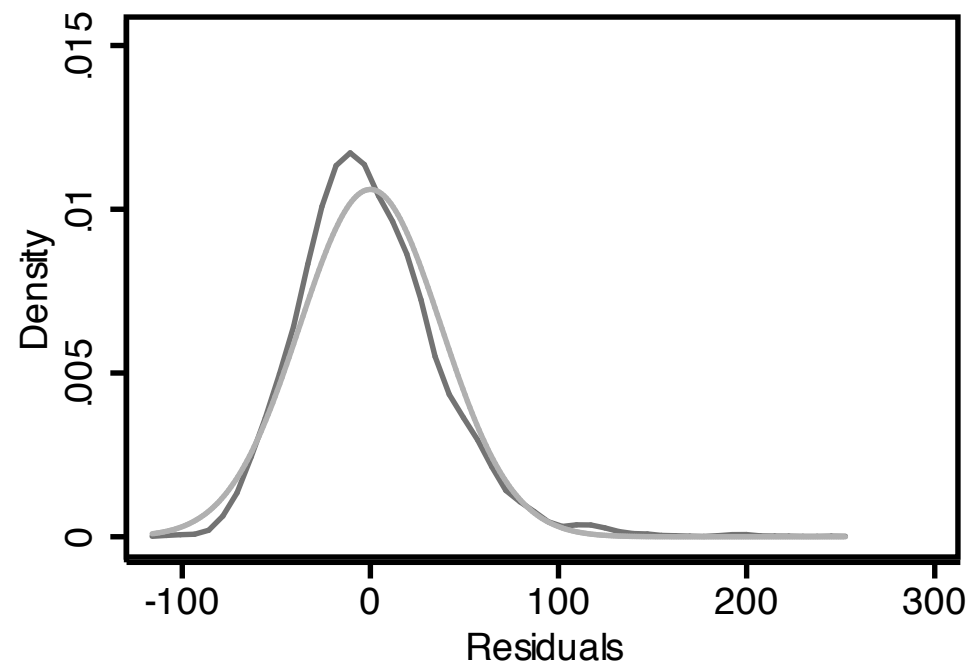
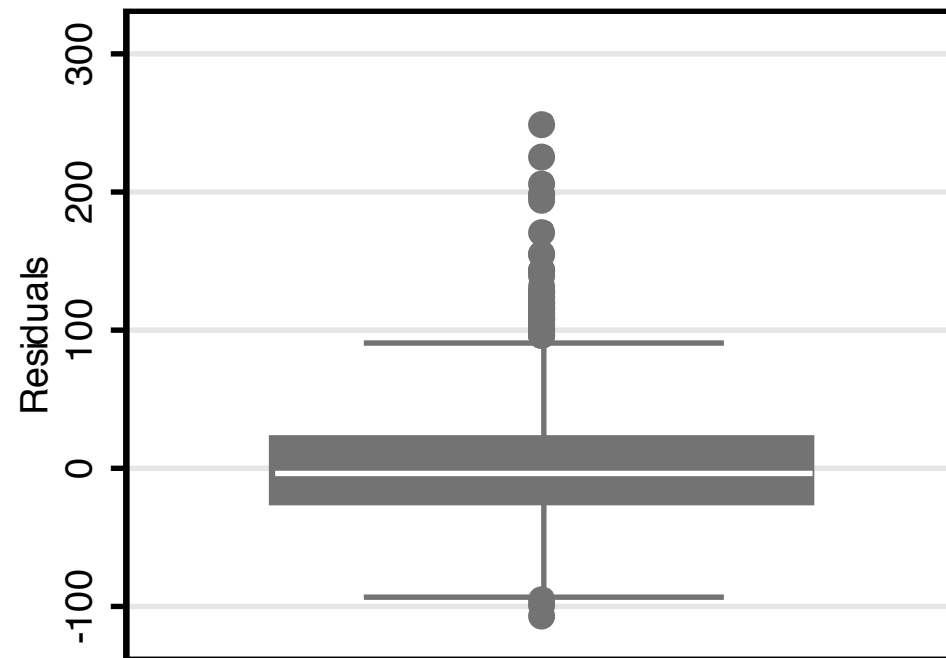
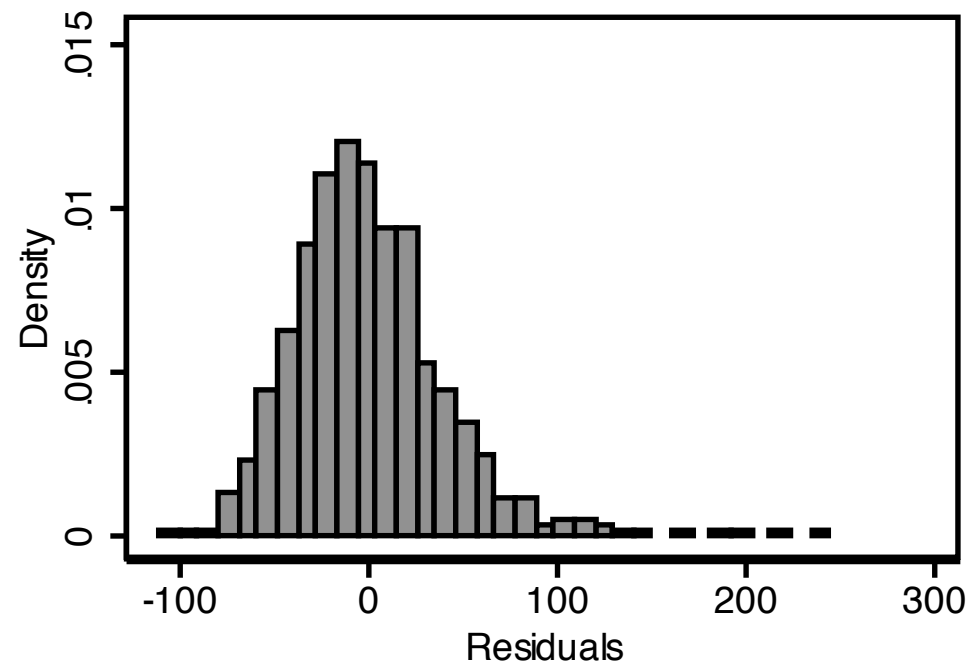
Source	SS	df	MS	Number of obs	=	2745
Model	42279.1877	5	8455.83753	F(5, 2739)	=	5.97
Residual	3881903.3	2739	1417.27028	Prob > F	=	0.0000
				R-squared	=	0.0108
				Adj R-squared	=	0.0090
Total	3924182.49	2744	1430.09566	Root MSE	=	37.647

LDL	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BMI	.3591038	.1341047	2.68	0.007	.0961472	.6220605
age	-.1897166	.1130776	-1.68	0.094	-.4114427	.0320095
nonwhite	5.219436	2.323673	2.25	0.025	.6631079	9.775764
smoking	4.750738	2.210391	2.15	0.032	.4165362	9.08494
drinkany	-2.722354	1.498854	-1.82	0.069	-5.661352	.2166445
_cons	147.3153	9.256449	15.91	0.000	129.165	165.4656

```
* save residuals in new variable
```

```
. predict resid, residuals
```

Diagnosing departures from normality

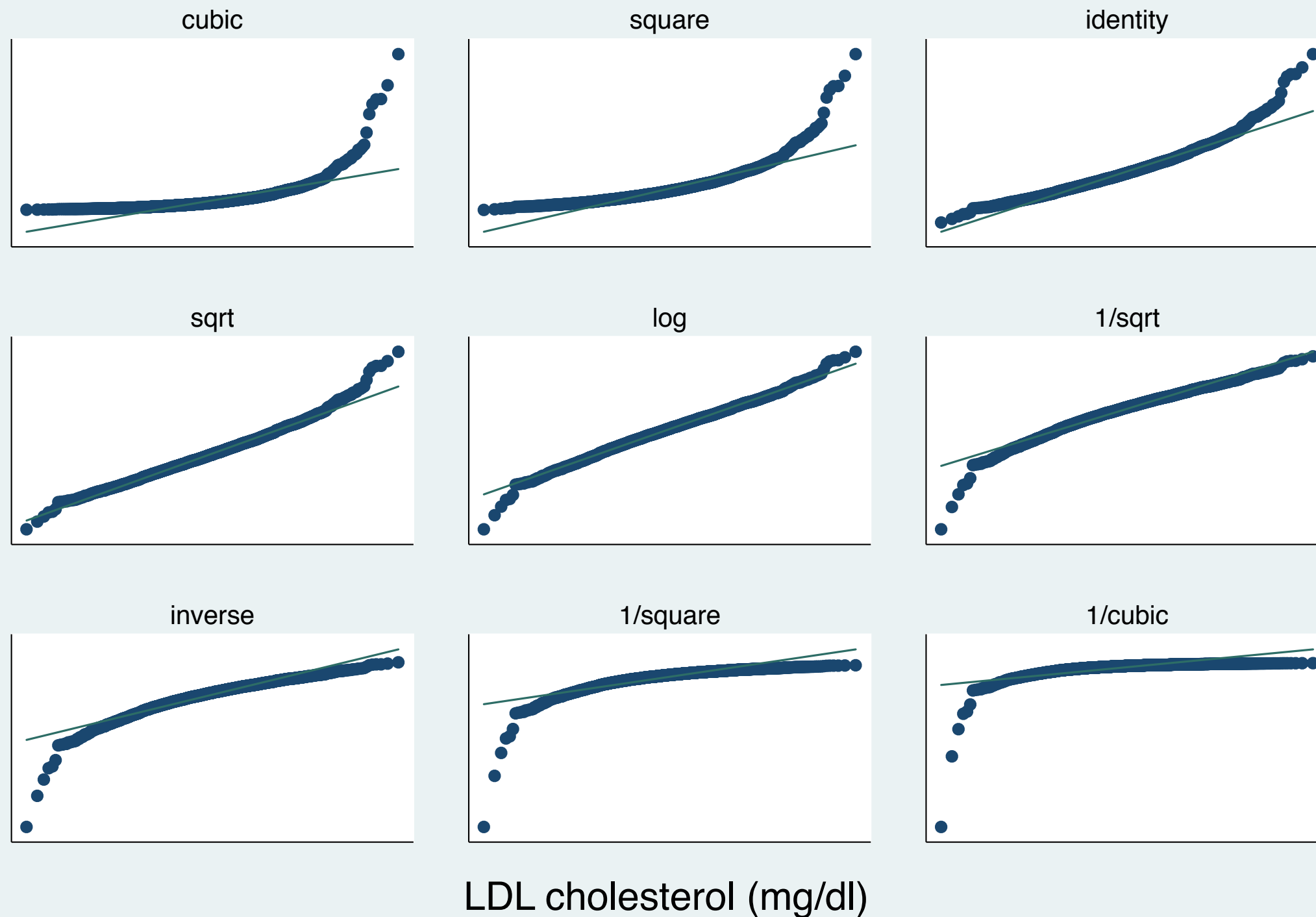


Some apparent evidence for skewness

Solution: transform the outcome

- Residuals skewed (usually to the right):
 - log, square root, other power transformations may help
 - may need to add constant for logs to make all values positive
- Search for best transformation using `q1adder` command in Stata
- Residuals symmetrically long-tailed
 - rank transformation, trimming, Winsorization

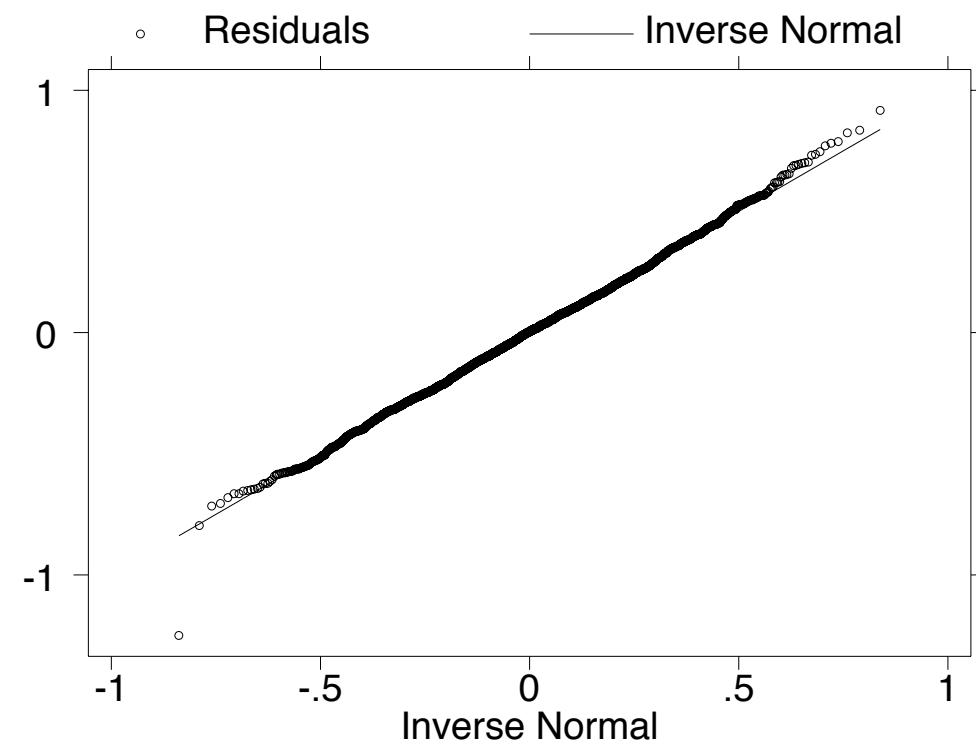
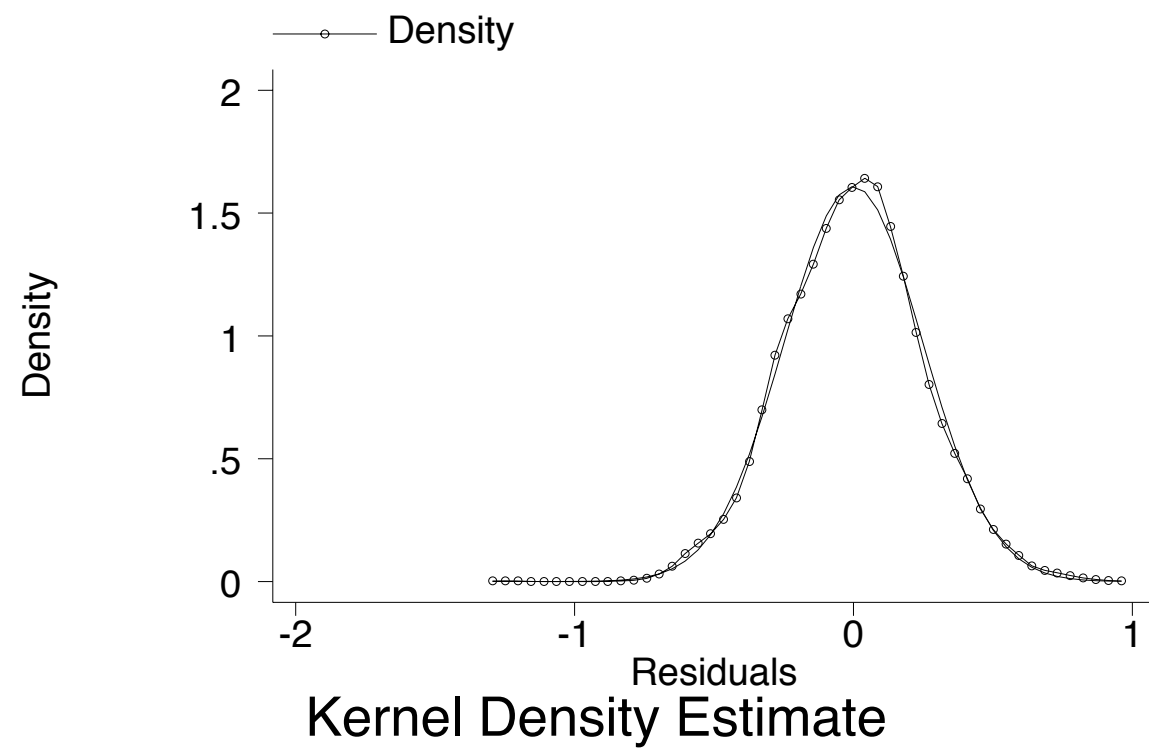
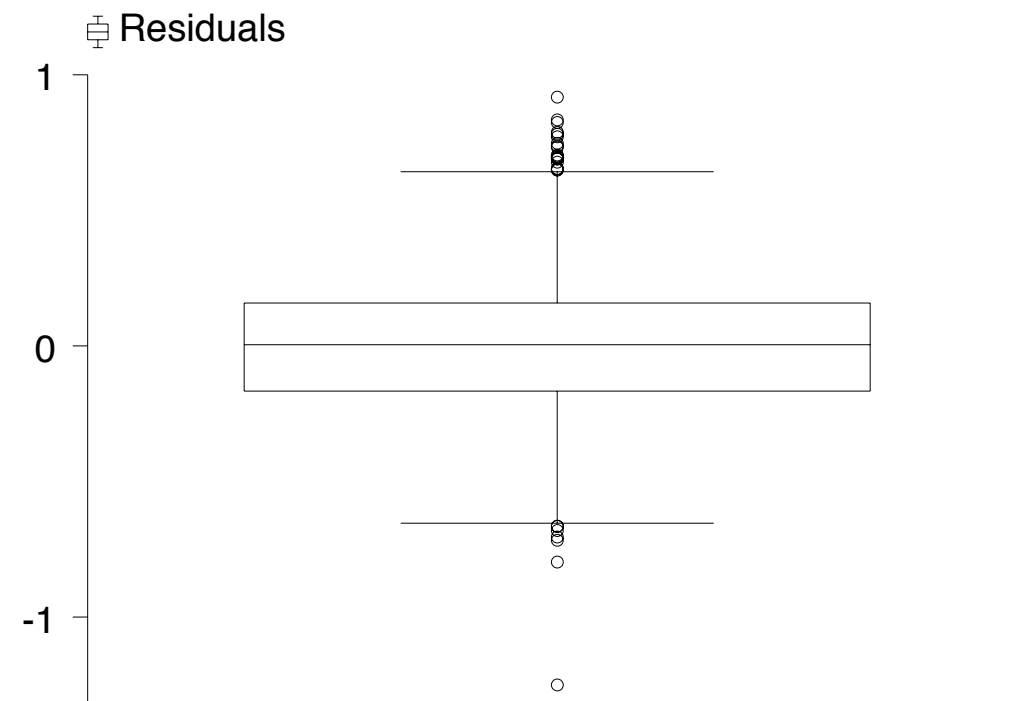
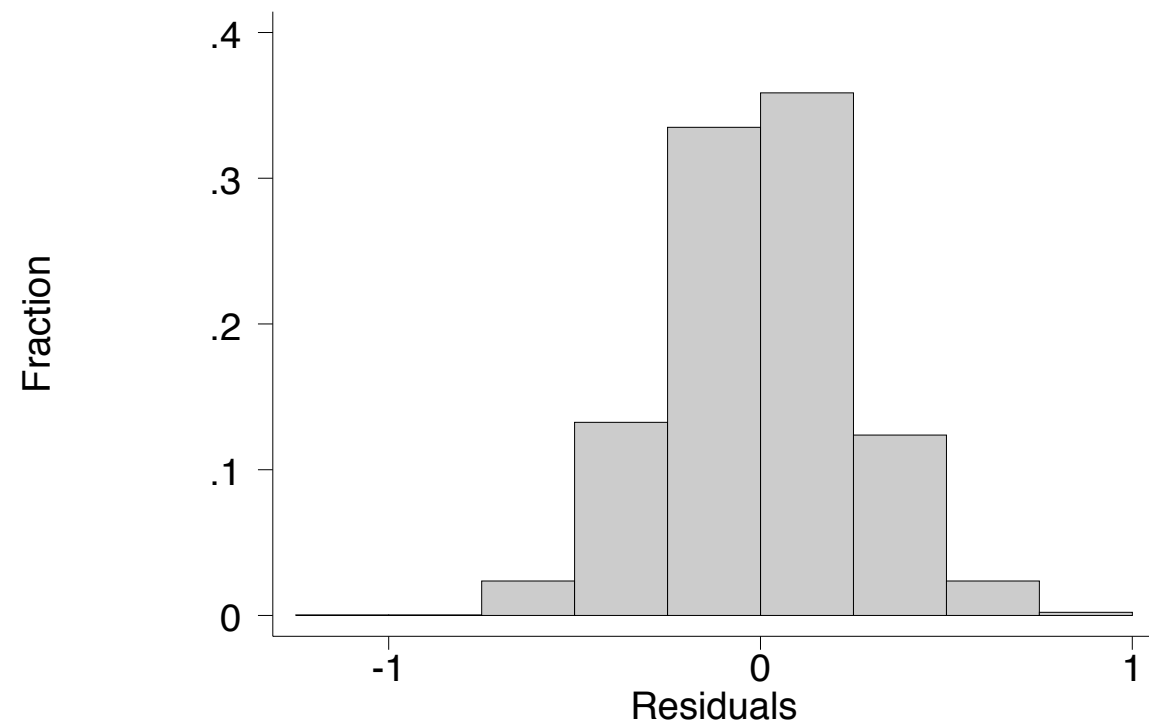
Q ladder plots for LDL



Quantile-Normal plots by transformation

Stata command: `qladder LDL, ylabel(none) xlabel(none)`

Residuals of log-transformed LDL



Another solution: bootstrap CIs

- Resample n observations *with replacement* from data, re-fit model, store estimates, repeat 100, 500, 1,000 times or more
- Distribution of bootstrap estimates represents sampling distribution of actual estimate
- Quick, partial solution (Normal-based bootstrap CIs):
 1. replace model-based SE by SD of bootstrap estimates
 2. construct CIs assuming Normality (i.e. estimate $\pm 1.96 \cdot \text{SE}$)

A better solution: percentile bootstrap CIs

- 95% CI: 2.5th to 97.5th percentile of bootstrap estimates
- Bias-correction shifts CI slightly to right or left
 - corrects for bias in the distribution of bootstrap estimates
- Slower but avoids making Normality assumption
- Requires using many (preferably ≥ 1000) bootstrap samples
 - extreme percentiles are noisy!
 - Stats's default of 50 repetitions *too small*

Stata default code for normal based bootstrap CIs

```
. regress LDL BMI age nonwhite smoking drinkany, vce(bootstrap)
(running regress on estimation sample)
```

```
Bootstrap replications (50)
-----+----- 1 -----+----- 2 -----+----- 3 -----+----- 4 -----+----- 5
..... 50
```

```
Linear regression                Number of obs      =      2745
                                Replications         =       50
                                Wald chi2(5)          =     28.49
                                Prob > chi2           =     0.0000
                                R-squared              =     0.0108
                                Adj R-squared          =     0.0090
                                Root MSE            =    37.6467
```

-----+-----							
	LDL	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
-----+-----							
	BMI	.3591038	.1462864	2.45	0.014	.0723878	.6458199
	age	-.1897166	.1020424	-1.86	0.063	-.3897161	.0102829
	nonwhite	5.219436	2.638803	1.98	0.048	.047478	10.39139
	smoking	4.750738	2.186506	2.17	0.030	.465265	9.036211
	drinkany	-2.722354	1.540028	-1.77	0.077	-5.740754	.2960465
	_cons	147.3153	8.558059	17.21	0.000	130.5418	164.0888
-----+-----							

Stata code for bias-corrected percentile CIs

```
. quietly regress LDL BMI age nonwhite smoking drinkany, vce(bootstrap, reps(1000))  
. estat bootstrap
```

Linear regression

Number of obs = 2745
Replications = 1000

LDL	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf. Interval]		
BMI	.35910384	.0041909	.13067908	.1089804	.6241642	(BC)
age	-.1897166	-.0004125	.11328262	-.3846027	.0482868	(BC)
nonwhite	5.2194359	.0693828	2.5754791	.0820059	10.25661	(BC)
smoking	4.750738	.0231322	2.2627561	.4212856	9.389248	(BC)
drinkany	-2.7223535	.041779	1.5069455	-5.879202	-.0033287	(BC)
_cons	147.31529	-.1740116	9.0724581	129.3511	164.1917	(BC)

(BC) bias-corrected confidence interval

Note:“estat bootstrap, all“ provides normal, percentile & BC intervals

Solution: Generalized Linear Models (GLMs)

- GLMs are available for a variety of outcome distributions and are characterized by a *link function* for the relationship between mean and predictors
- Example: logistic model for binary outcomes uses the log odds to link mean (proportion) and predictor:

$$\log \frac{E[y|x_1, x_2, \dots]}{1 - E[y|x_1, x_2, \dots]} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

- Other generalized linear models (GLMs) for various (e.g. count and continuous) outcomes, frequently model $\log E[y | x]$ as linear (Biostat 209)

<i>outcome type</i>	<i>distribution</i>	<i>variance-mean relationship</i>	<i>default link</i>
continuous	normal	constant	identity
continuous	gamma	proportional ($\sigma^2 \propto \mu$)	inverse
binary	binomial	$\sigma^2 = n \mu(1 - \mu)$	log odds
count	Poisson	equal ($\sigma^2 = \mu$)	log
count	neg. binomial	$\sigma^2 = \mu + \mu^2/k$	log

Another solution: ordinal categorical models

- Examples:
 - ─ distribution of Agatston scores for coronary artery calcium (CAC)
mostly zeroes with long right tail
- Log-transformation (after adding 1) does not help: still mostly zeroes with long right tail
- Could dichotomize outcome as $CAC > 0$ or $CAC > 10$
 - ─ use logistic model - but potentially wasteful

Ordinal categorical models (continued)

- Alternatively, categorize CAC as 0, 1-9, 10-99, 100-399, ≥ 400 , use regression model for ordinal outcomes
- proportional odds (`ologit` - to be discussed with logistic regression)

Solution: two-part and zero-inflated models

- *Two-part models*:
 1. logistic model for $CAC > 0$
 2. linear model for log-transformed CAC scores > 0
- Testing of overall predictor effects uses *seemingly unrelated regressions* approach, `suest` command
- *Zero-inflated models* allow for more zeroes than expected under Poisson or negative binomial distribution (and can be applied to non-count outcomes); `zip` and `zinb` commands

Checking normality: summary

- Diagnostics: curvature in QQ-plot of residuals
- Solutions: transform outcome, use bootstrap percentile CIs, or GLMs including ordinal, two-part, zero-inflated models

Model checking: constant variance

- Even Normally distributed residuals may be characterized by unequal variance (aka *heteroskedasticity*)
- If constant variance assumption is violated
 - coefficient estimates unbiased but inefficient (i.e. standard errors may be overestimated)
 - tests for between-group differences may be invalid

Diagnostics: constant variance

- Plot residuals against fitted values, predictors
 - ─ check for horizontal funnel shapes
- Compare sample size, variance of residuals across subgroups:
 - ─ watch out if both differ by factors of more than 2

Solution: transform outcome

Some standard variance-stabilizing transformations:

<i>outcome</i>	<i>transformation</i>
variance $\propto$ mean	square root
SD $\propto$ mean	log
proportions	arcsin
correlations	$\log[(1 + \rho)/(1 - \rho)]$

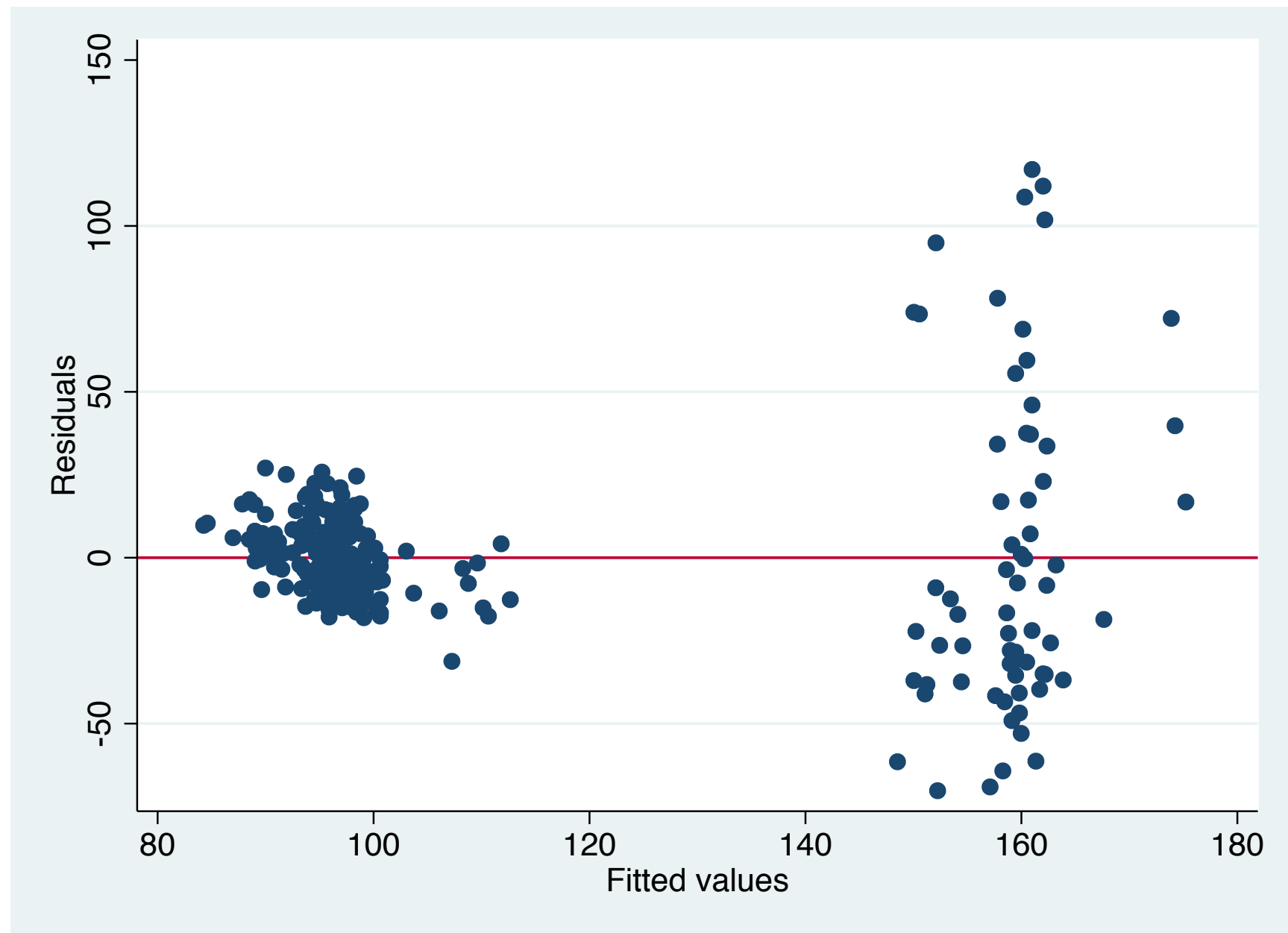
$\propto$ = “proportional to”

ρ = correlation coefficient

Example: model for mean glucose HERS

```
*  
. use hersdata, clear  
* take a 10% sample to make visualization easier (for illustration)  
. sample 10  
(2487 observations deleted)  
* untransformed outcome - fit model & plot residuals  
. quietly regress glucose i.diabetes i.physact age i.raceth i.smoking i.drinkany  
. predict resid1, residuals  
. rvfplot, yline(0)
```

RVF plot - untransformed outcome



Comparing residual variance by subgroup (un-transformed outcome)

```
. tabstat resid1, by(physact) stat(n var) nototal
```

Summary for variables: resid1

by categories of: physact (comparative physical activity)

physact	N	variance
much less active	26	1057.983
somewhat less ac	46	792.155
about as active	87	731.2216
somewhat more ac	85	554.0864
much more active	32	107.7953

```
. tabstat resid1, by(diabetes) stat(n var) nototal
```

Summary for variables: resid1

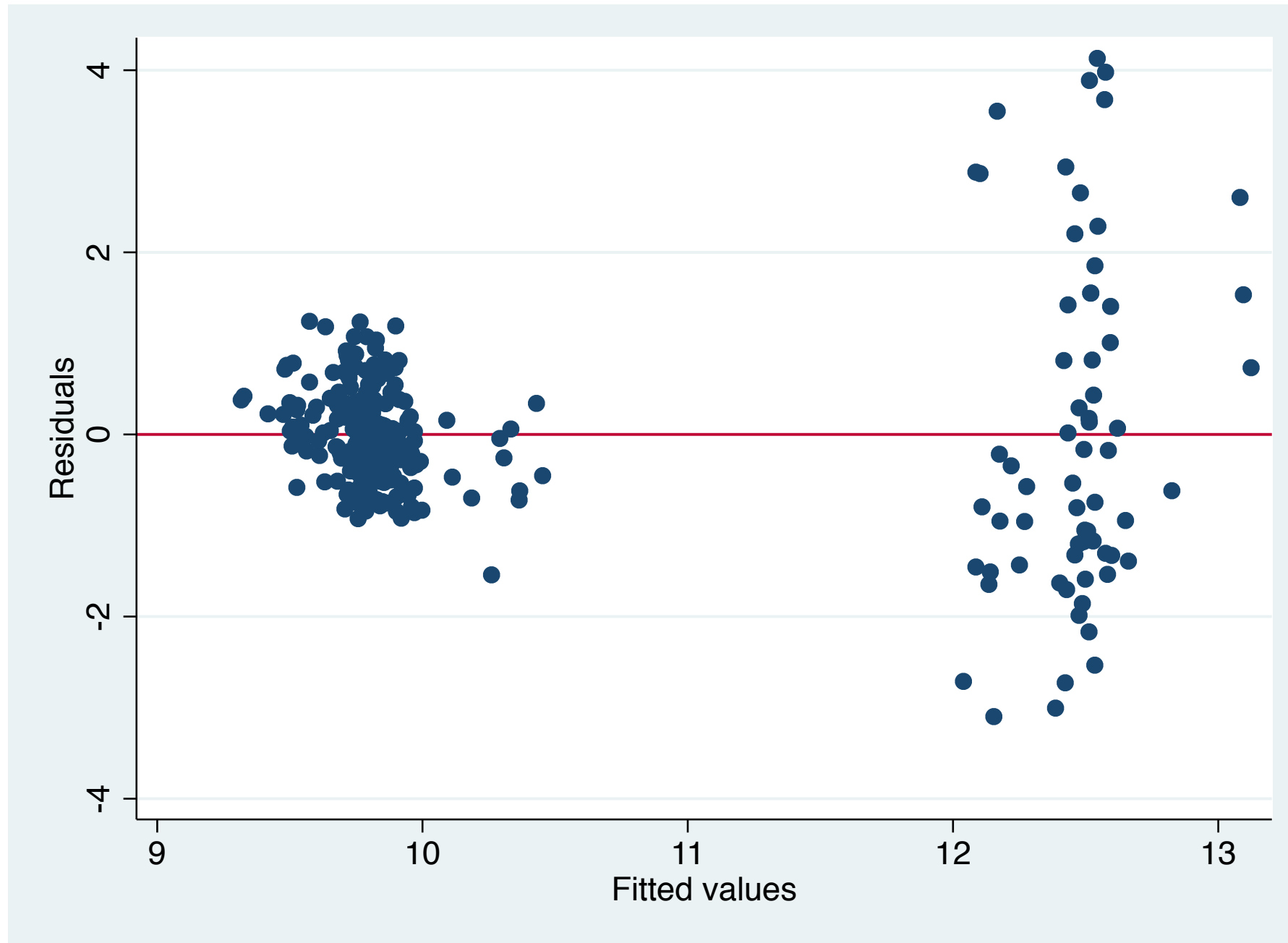
by categories of: diabetes (diabetes)

diabetes	N	variance
no	196	30.45264
yes	80	53.23272

Example: model for mean glucose HERS - square root transformed

```
* square root transformed outcome - fit model & plot residuals
. gen sqrt_glucose = sqrt(glucose)
. quietly regress sqrt_glucose diabetes i.physact age i.raceth smoking drinkany
. rvfplot, yline(0)
. graph export ./figures/rvp2.pdf, replace
```


RVF plot - square root transformed outcome



Transformation not effective in this case

Solution: use robust SEs

```
. regress glucose diabetes i.physact age i.raceth smoking drinkany, vce(robust)
. . .
```

glucose		Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
diabetes		55.32816	5.704065	9.70	0.000	44.09711	66.55922
physact							
somewhat less active		.5986391	7.670311	0.08	0.938	-14.50387	15.70115
about as active		6.51184	7.519767	0.87	0.387	-8.294252	21.31793
somewhat more active		2.873804	7.282648	0.39	0.693	-11.46541	17.21302
much more active		.4625191	6.907942	0.07	0.947	-13.13892	14.06396
age		-.3130465	.2466262	-1.27	0.205	-.7986428	.1725497
raceth							
African American		9.907849	7.805314	1.27	0.205	-5.460473	25.27617
Other		22.48085	15.08384	1.49	0.137	-7.218569	52.18027
smoking		-4.696382	4.223875	-1.11	0.267	-13.01301	3.620243
drinkany		6.649252	3.427625	1.94	0.053	-.0995925	13.3981
_cons		112.8064	16.89753	6.68	0.000	79.53592	146.0769

Note: robust standard errors are estimated *without* assuming a distribution for the residuals, and will generally be larger than the standard errors based on normally distributed residuals

... or use more conservative robust SEs

```
. regress glucose diabetes i.physact age i.raceth smoking drinkany, vce(hc3)
. . .
```

glucose		Robust HC3		t	P> t	[95% Conf. Interval]	
		Coef.	Std. Err.				
	diabetes	55.32816	5.838609	9.48	0.000	43.8322	66.82413
	physact						
somewhat less active		.5986391	8.082405	0.07	0.941	-15.31526	16.51254
about as active		6.51184	7.877636	0.83	0.409	-8.998881	22.02256
somewhat more active		2.873804	7.619965	0.38	0.706	-12.12957	17.87718
much more active		.4625191	7.247594	0.06	0.949	-13.80768	14.73271
	age	-.3130465	.2557038	-1.22	0.222	-.8165162	.1904231
	raceth						
African American		9.907849	8.189902	1.21	0.227	-6.21771	26.03341
Other		22.48085	16.98321	1.32	0.187	-10.95835	55.92005
	smoking	-4.696382	4.444625	-1.06	0.292	-13.44765	4.054891
	drinkany	6.649252	3.505505	1.90	0.059	-.2529339	13.55144
	_cons	112.8064	17.51732	6.44	0.000	78.31558	147.2972

- *Note:* hc3 option provides more conservative robust SE's that may perform better for models with pronounced non-constant variance and/or smaller samples sizes (compare SEs and CIs with previous slide)

Solution: Generalized linear model (GLM)

- GLMs model relationship between mean and predictors
 - mean linked to predictors via a *link function*
 - include a variety of outcome distributions (each characterized by a variance-mean relationship)
- Choose a GLM with the appropriate variance to mean relationship:

<i>outcome type</i>	<i>distribution</i>	<i>variance-mean relationship</i>	<i>std. link</i>
continuous	normal	constant	identity
continuous	gamma	proportional ($\sigma^2 \propto \mu$)	inverse
binary	binomial	$\sigma^2 = n \mu(1 - \mu)$	log odds
count	Poisson	equal ($\sigma^2 = \mu$)	log
count	neg. binomial	$\sigma^2 = \mu + \mu^2/k$	log

Note: GLMs will be covered in Biostat 209 (also, see table 8.8 in text)

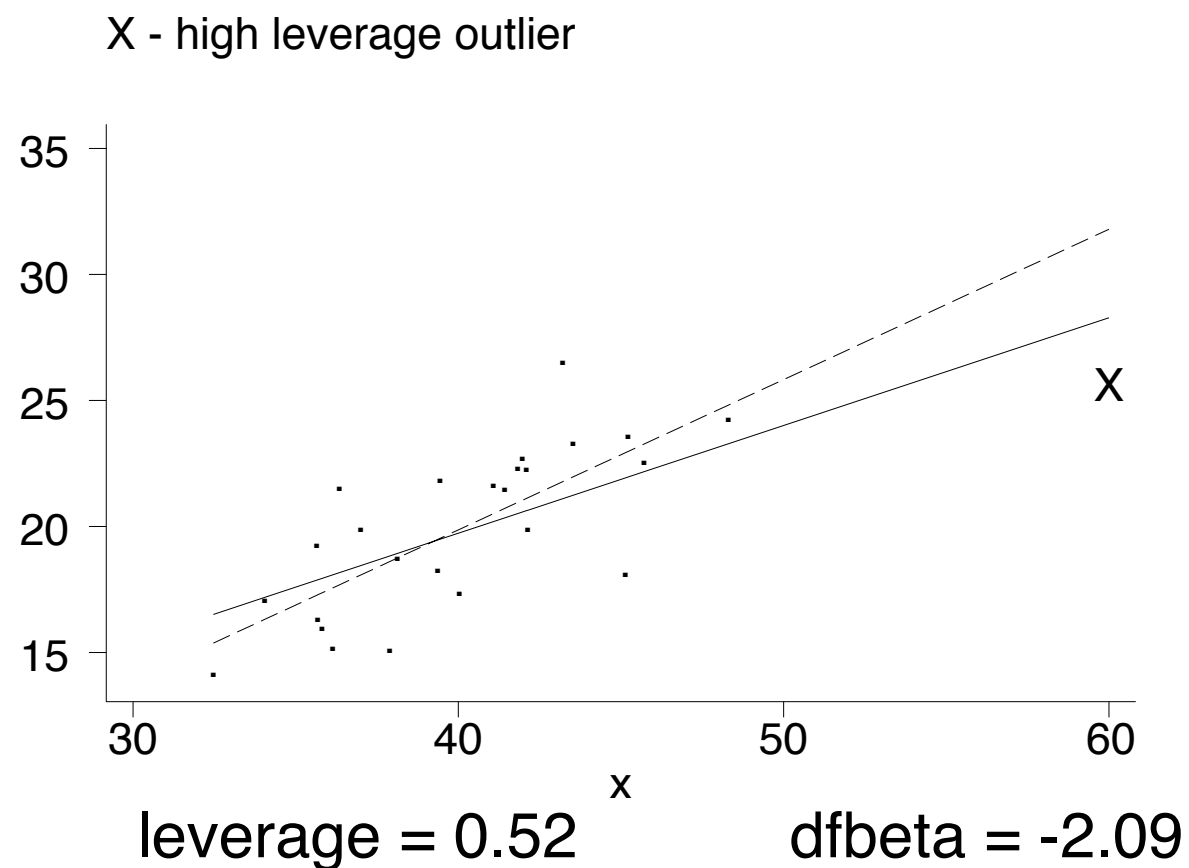
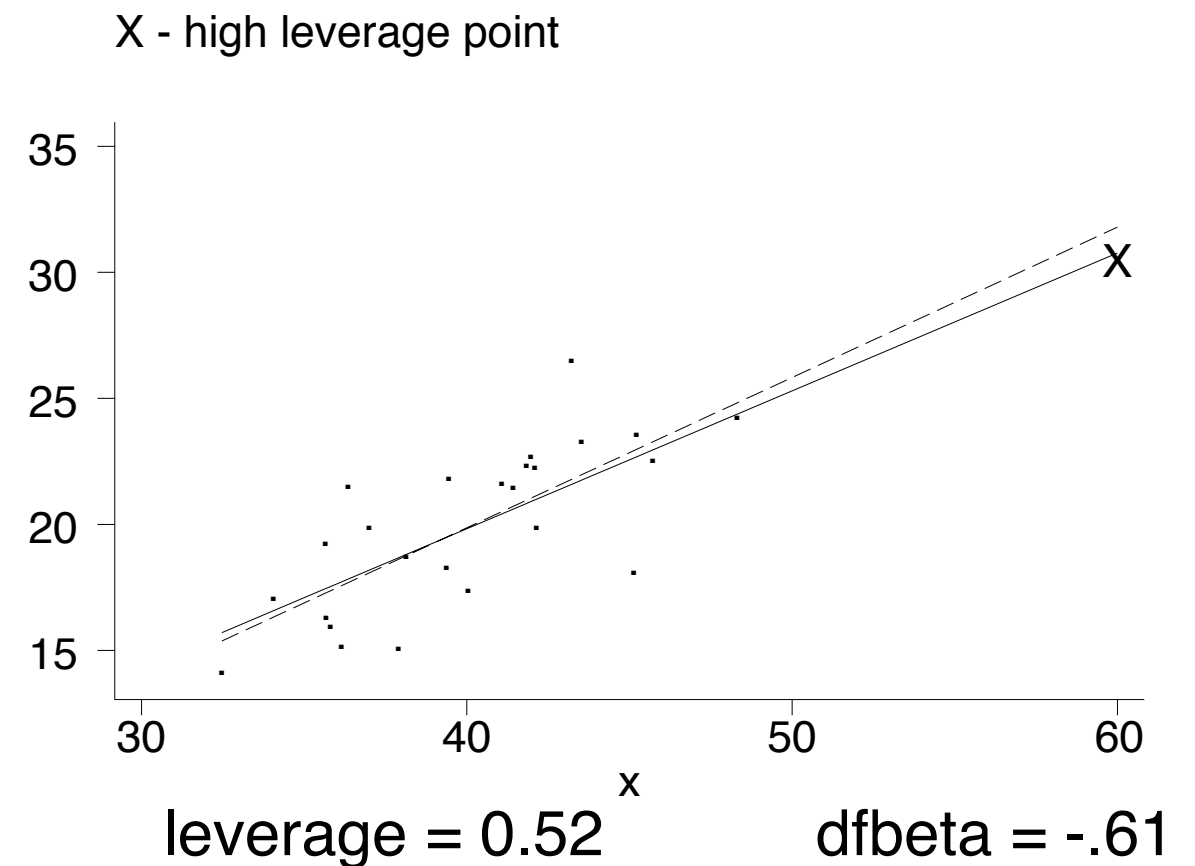
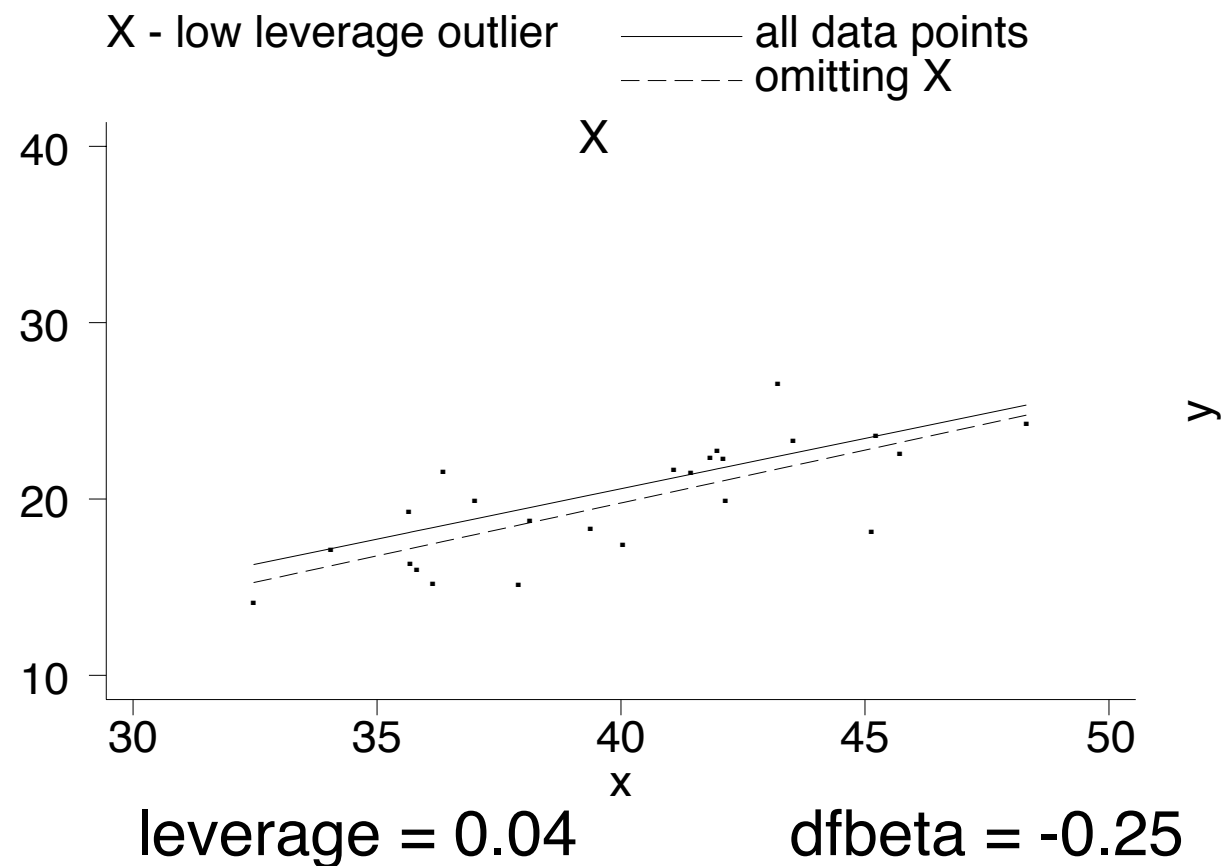
Checking constant variance: summary

- Diagnostics: funnel shapes in RVP plot, variable n , SD across subgroups
- Solutions: transform outcome, use robust SEs or GLMs

Model checking: high leverage and influential points

- *High-leverage:*
 - ≥ 1 extreme predictor, or anomalous combination of predictors
 - potential to influence coefficient estimates unduly
- *Influential:*
 - high-leverage plus big impact on coefficients
- Inferences that are highly dependent on a few influential observations are potentially misleading

Examples: Simple outlier, high leverage, high influence



Diagnostics: boxplots of dfbeta statistics

- dfbeta statistics measure changes in each β_j when each data point is omitted
- Defined for each observation and predictor in model
- dfbeta command is issued after model is fitted (results in a dfbeta measure for each predictor in model):

```
. quietly regress LDL BMI age nonwhite smoking drinkany
```

```
. dfbeta
```

```
(18 missing values generated)
```

```
    _dfbeta_1: dfbeta(BMI)
```

```
(18 missing values generated)
```

```
    _dfbeta_2: dfbeta(age)
```

```
(18 missing values generated)
```

```
    _dfbeta_3: dfbeta(nonwhite)
```

```
(18 missing values generated)
```

```
    _dfbeta_4: dfbeta(smoking)
```

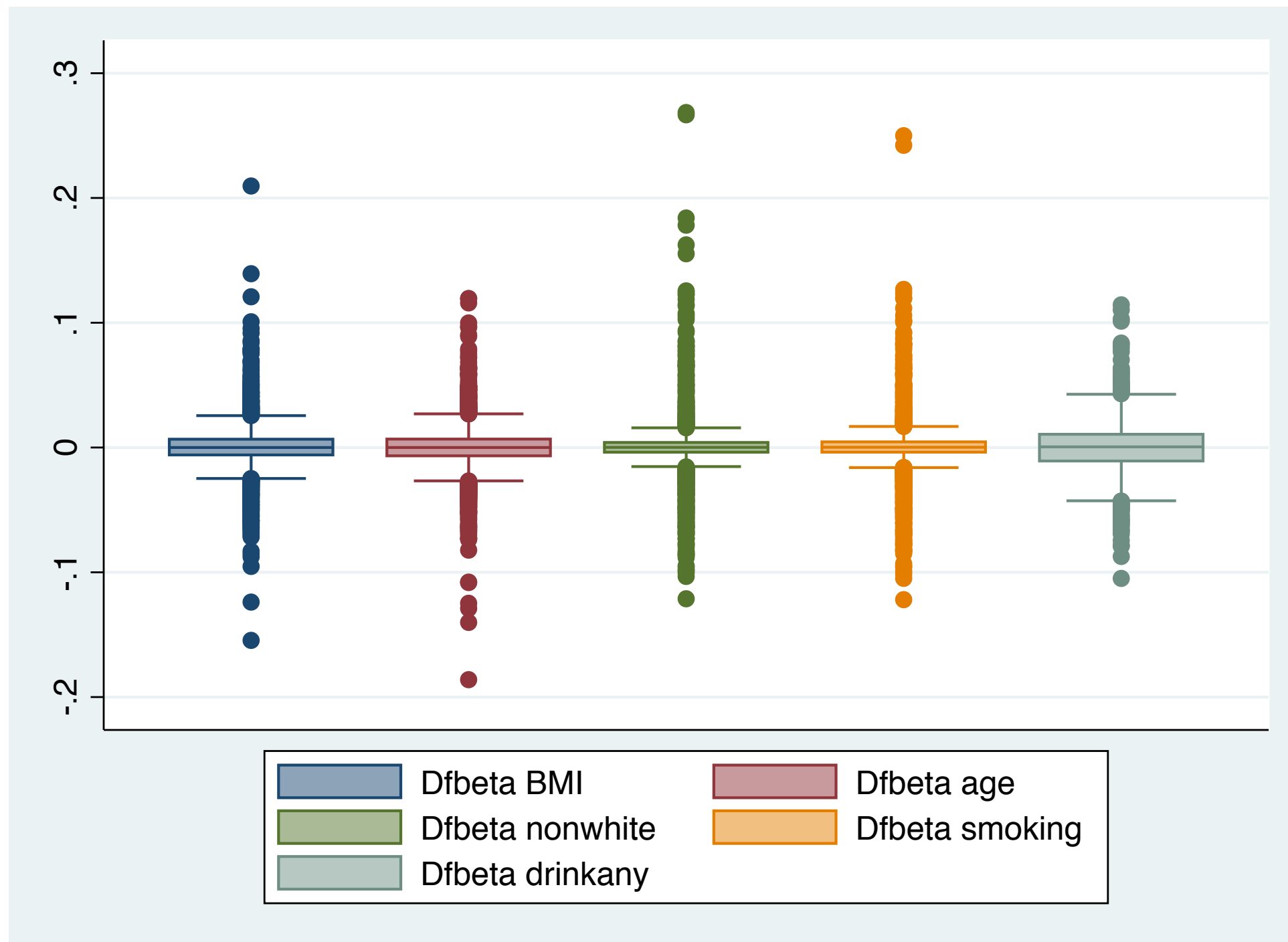
```
(18 missing values generated)
```

```
    _dfbeta_5: dfbeta(drinkany)
```

- Boxplots of dfbetas can be used to check for outliers

```
. graph box _dfbeta_1 _dfbeta_2 _dfbeta_3 _dfbeta_4 _dfbeta_5
```


Boxplots of dfbetas for BMI - LDL model (slide 27)



Diagnostics: manual checking dfbeta statistics:

- Identify up to 10 observations with biggest dfbeta values

* Stata statements to list observations with large dfbeta (yields four observations)

* Statement for first predictor (BMI) below (repeat for other predictors)

```
. list id BMI age nonwhite smoking drinkany _dfbeta_1 if abs(_dfbeta_1) > 0.2 & !=missing(_dfbeta_1)
```

```
+-----+
|   id   BMI   age  nonwhite  smoking  drinkany  _dfbet~1 |
+-----+
70. |    70   28.74   63         yes       no        no   .2666293 |
1768. |  1768   28.22   61         yes       yes        no   .2684654 |
1020. |  1020   26.22   58         no        yes       yes   .2422048 |
1768. |  1768   28.22   61         yes       yes        no   .2498547 |
+-----+
```

- Check for data errors or other anomaly
- Re-fit model without influential points, reassess conclusions, report sensitivities
- Consider deleting influential points if they represent errors, or individuals that don't represent the target population

Sensitivity of LDL model to 4 influential points with $df\beta a > 0.2$ in absolute value

<i>predictor</i>	<i>all observations</i>		<i>omitting 4 points</i>	
	β	p	β	p
BMI	0.36	0.0007	0.34	0.010
Age	-1.89	0.090	-1.86	0.090
Nonwhite	5.22	0.025	4.19	0.066
Smoking	4.75	0.032	3.78	0.072
Alcohol use	-2.72	0.069	-2.64	0.072

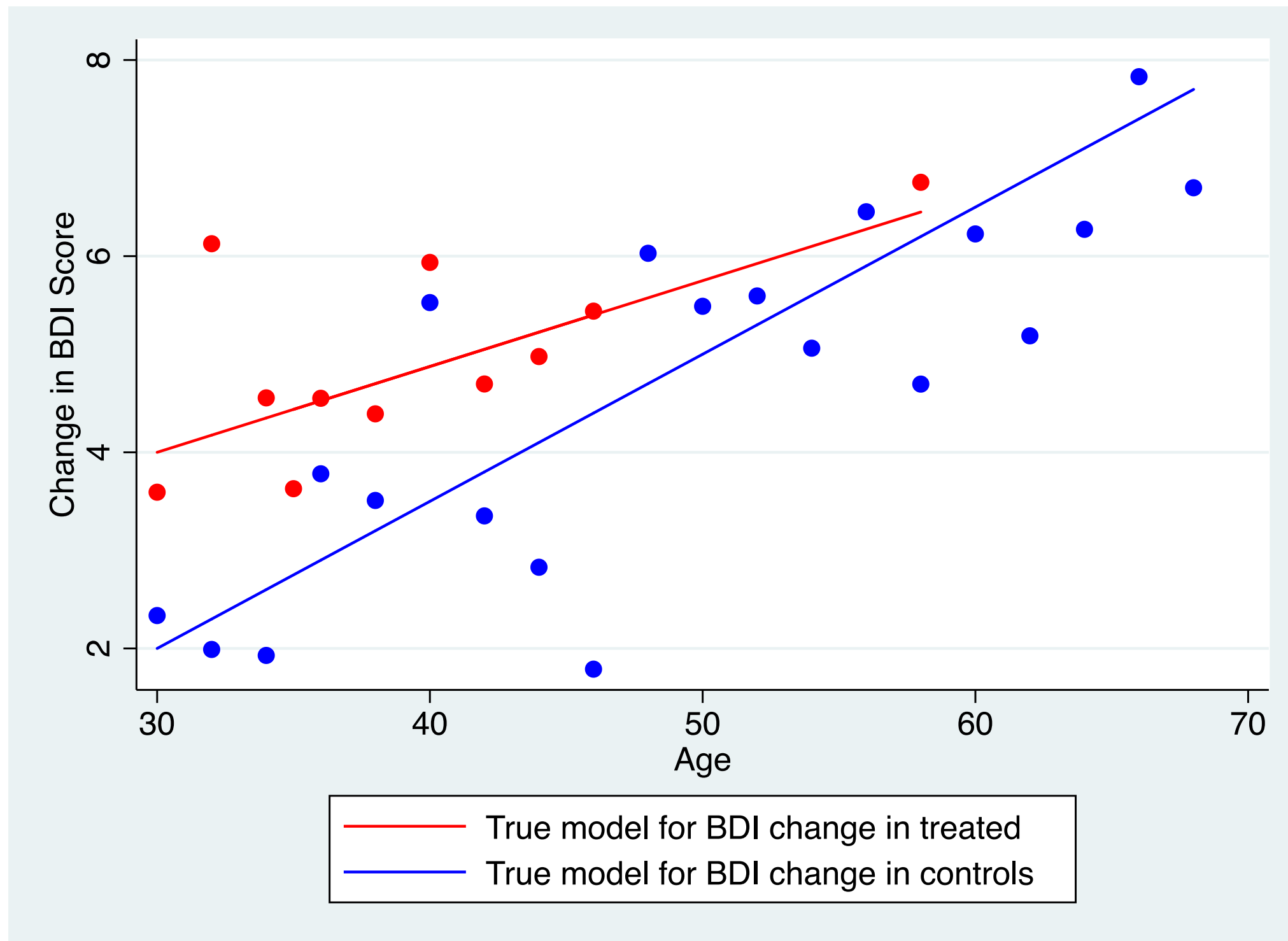
Checking influential points: summary

- Diagnostics: boxplots of $dfbetas$
- Solutions: fix errors, conduct sensitivity analyses omitting influential points

Model checking: predictor/covariate overlap

- Regression evaluation of binary exposure/treatment may be unreliable if exposed, unexposed differ (i.e. don't *overlap*) with respect to other predictors
- Lack of overlap effects:
 - may lead to biased coefficient estimates, especially in small datasets
 - reduced power to detect between-group differences
 - conclusions about effect of exposure may be based on *extrapolation*

Example: Lack of age overlap in model for effect of treatment on Beck Depression Inventory score



Data points are from a single sample of 31 individuals (11 **treated**)

Regression model fitted to observed data ($n=31$)

- shows low power to detect treatment x age interaction
- information on treatment effects in ages > 50 is limited by non-overlap

```
. regress del_bdi i.treatment##c.age
```

Source	SS	df	MS	Number of obs = 31			
Model	46.3692007	3	15.4564002	F(3, 27) = 15.42			
Residual	27.0583639	27	1.00216163	Prob > F = 0.0000			
Total	73.4275647	30	2.44758549	R-squared = 0.6315			
				Adj R-squared = 0.5906			
				Root MSE = 1.0011			

del_bdi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.treatment	3.217112	1.88746	1.70	0.100	-.6556366	7.08986
age	.1247361	.0194101	6.43	0.000	.0849098	.1645623
treatment#c.age						
1	-.0429515	.0445653	-0.96	0.344	-.1343918	.0484889
_cons	-1.483581	.9770828	-1.52	0.141	-3.488389	.5212275

Solution: diagnosing lack of overlap

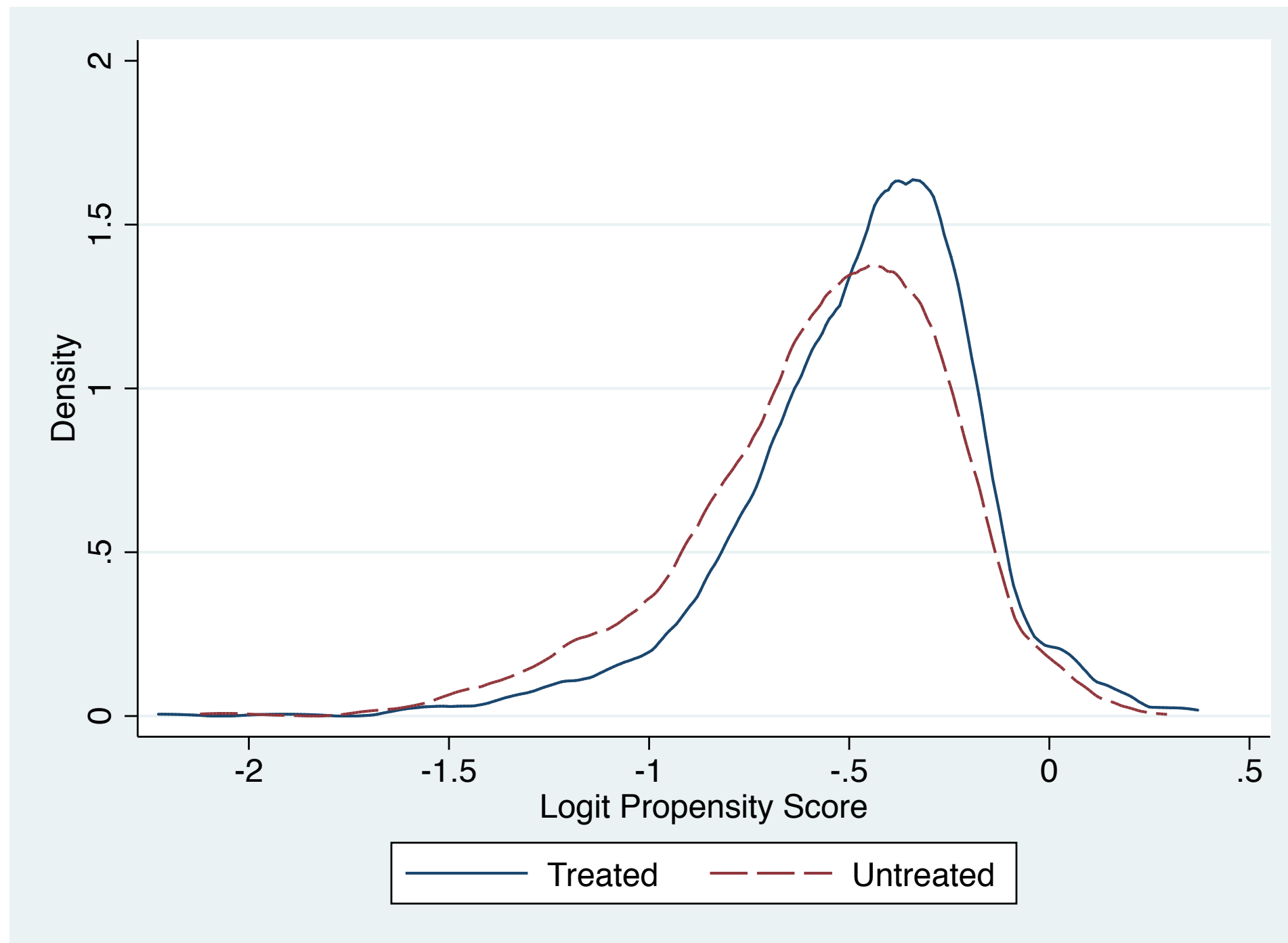
- Compare mean, quartiles, range of individual predictors in exposed and unexposed
 - Somewhat limited because lack of overlap may be manifested in *combinations* of predictors
- Use *propensity scores*
 - fit logistic model for primary predictor
 - ▶ include an MSAS (minim. suff. adjust. set) for the exposure-outcome relationship
 - ▶ capture non-linearities and interactions
 - ▶ results in estimated probabilities of exposure/treatment as a function of predictors in MSAS
 - get fitted values (on probability or linear predictor)
 - plot the results by primary predictor and check overlap

Example: Propensity score model for indicator of statin use in HERS

- *Goal:* assess relationship between statin use (Y/N) and a study outcome (e.g. CHD)
- Use propensity scores to investigate degree of overlap between users/non-users of statins

```
. * logistic model for statin use
. quietly logistic statins agesp* i.raceth i.educ_cat i.smoking##i.lessactive diabetes
. * calculate logit propensity score
. predict logit_ps, xb
. * density plots of logit scores in statin users and non-users
. twoway (kdensity logit_ps if statins==1, area(1) lpattern(solid)) ///
>         (kdensity logit_ps if statins==0, area(1) lpattern(longdash)), ///
>         ytitle("Density") xtitle("Logit Propensity Score") ///
>         legend(order(1 "Treated" 2 "Untreated")) ///
>         saving(pscores, replace)
```


Overlap diagnostics for statin use in HERS



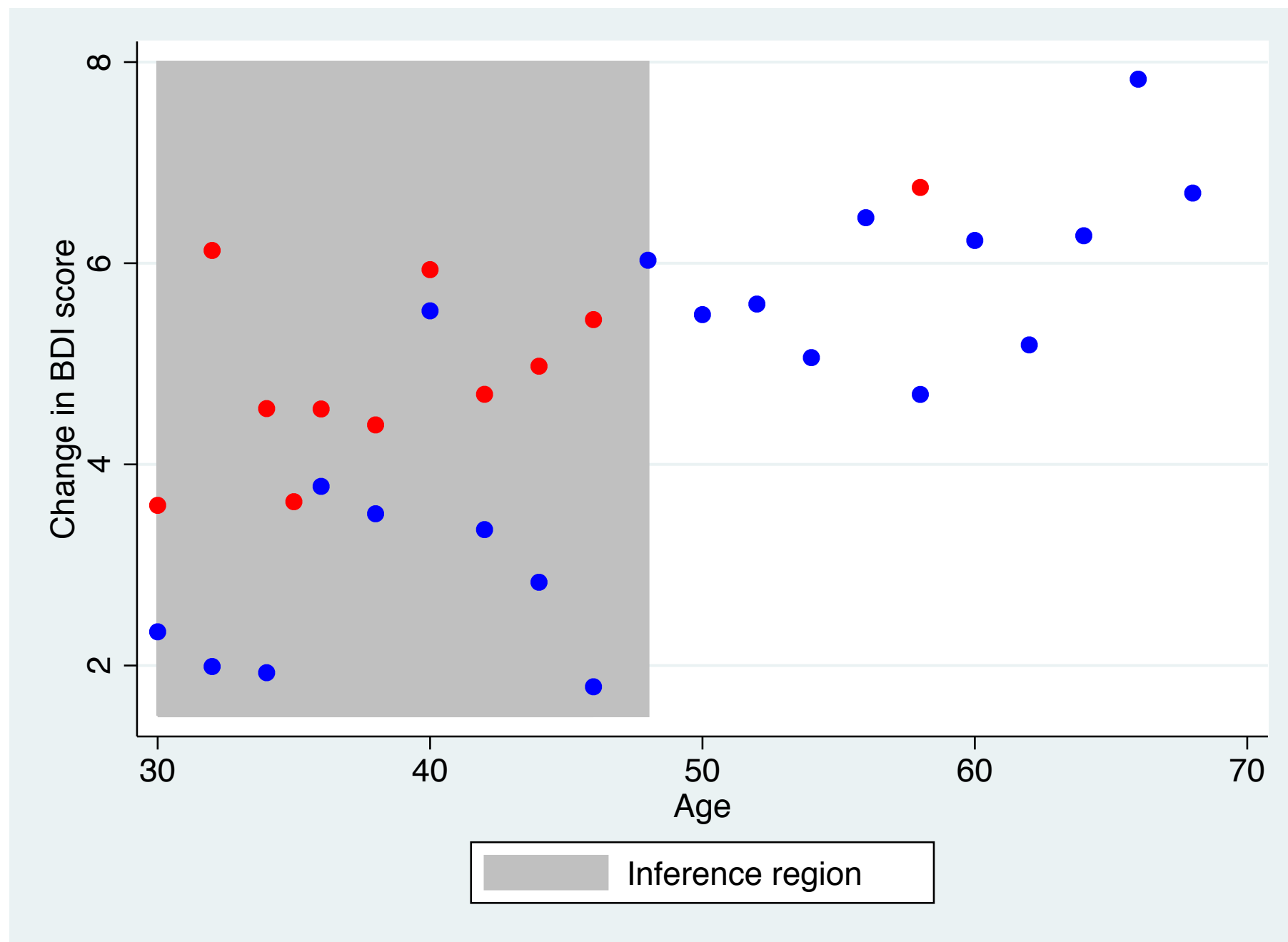
Overlap is generally quite good (thanks to treatment randomization)
- in this case no further action may be warranted

Solution: accounting for lack of overlap

When lack of overlap is an apparent problem:

- Restrict inference to region of good overlap between exposed/non-exposed
- Match on prognostic covariates or propensity scores (covered in Biostat 215)

Restricting inference to region of overlap (BDI example)



Compare conclusions based on observations with sufficient overlap to results from full sample

Model checking: summary

- Non-linearity:
 - Diagnostics: curved Lowess smooth in CPR or RVP plot
 - Solutions: transform predictor, including splines
- Non-normality:
 - Diagnostics: curvature in QQ-plot
 - Solutions: transform outcome, use bootstrap CIs, GLM or ordinal model

Model checking: summary

Transform or not?

- Transformations can help meet assumptions
 - ─ but make results harder to interpret
- If violations mild, results robust, reasonable not to transform
- If conclusions change substantially after transformation
 - ─ model that meets assumptions better is more reliable

Model checking: summary

- Non-constant variance:
 - Diagnostics: funnel shapes in RVP plot, SDs differ across unequal size subgroups
 - Solutions: transform outcome, use GLM, robust SEs
- Influential points:
 - Diagnostics: boxplots of $df\beta_a$ statistics
 - Solutions: identify up to 10 influential points, correct data errors, omit influential points if justifiable, present sensitivity analysis

Final note: independence assumption

The assumption of *independent outcomes* applies to all the regression models covered in this course.

- Dependence can arise in a variety of contexts:
 - longitudinal studies of recurrent outcomes (e.g. episodic diseases)
 - multiple “clustered” outcome measures at a single occasion (e.g. characteristics of teeth, disease occurrence within family members)

Regression methods that account for dependent outcomes will be covered in Biostat. 209.