Checking the model

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

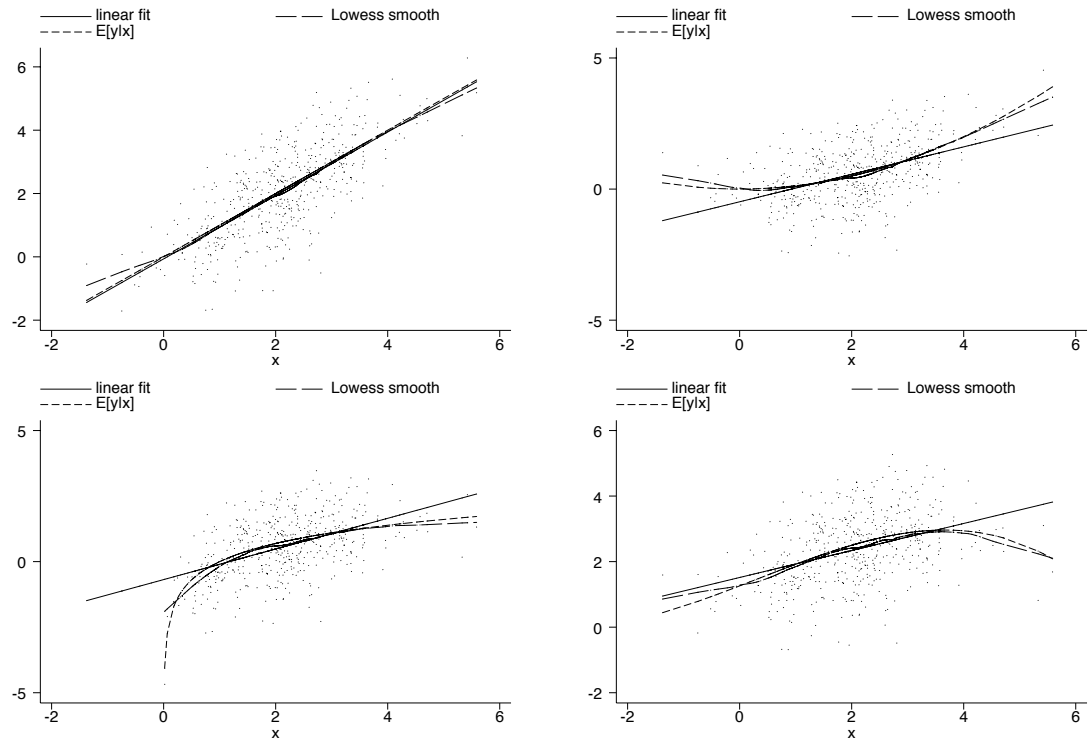
Checking the model: linearity

- Regression line assumed linear in continuous predictors
  - i.e., slope is constant
  - equivalently, has no curvature
- If model is correct
  - residuals have mean zero at every value of predictor
  - key to diagnostic tools for departures from linearity

Checking the model: 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: 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

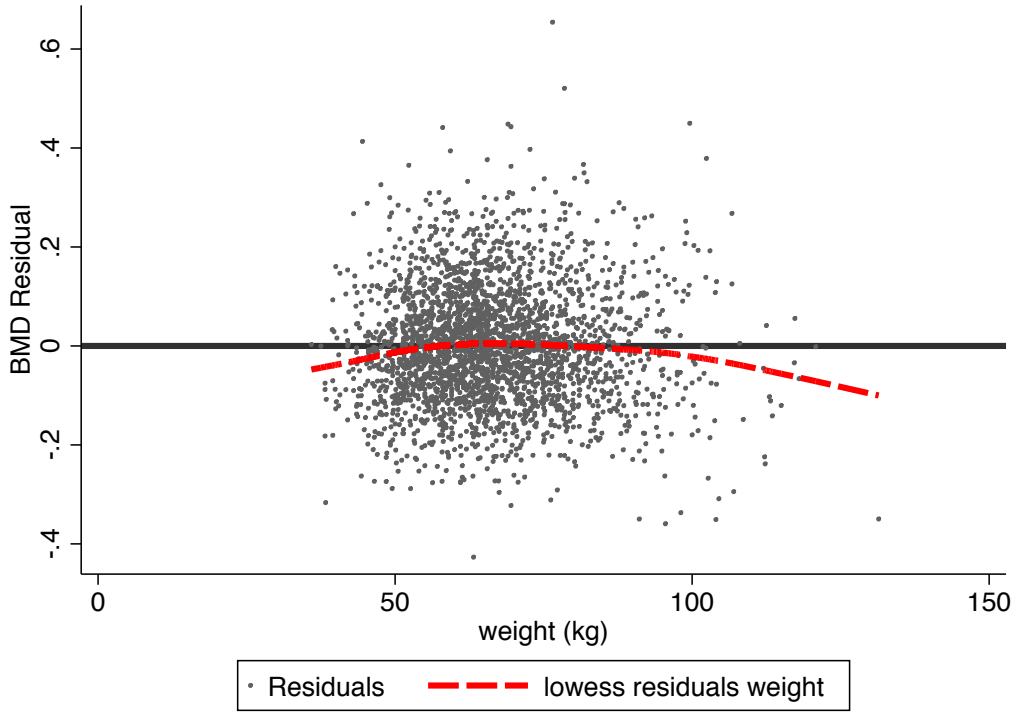
- Better alternative: component plus residual (CPR) plots

  – component due to predictor added back into residual

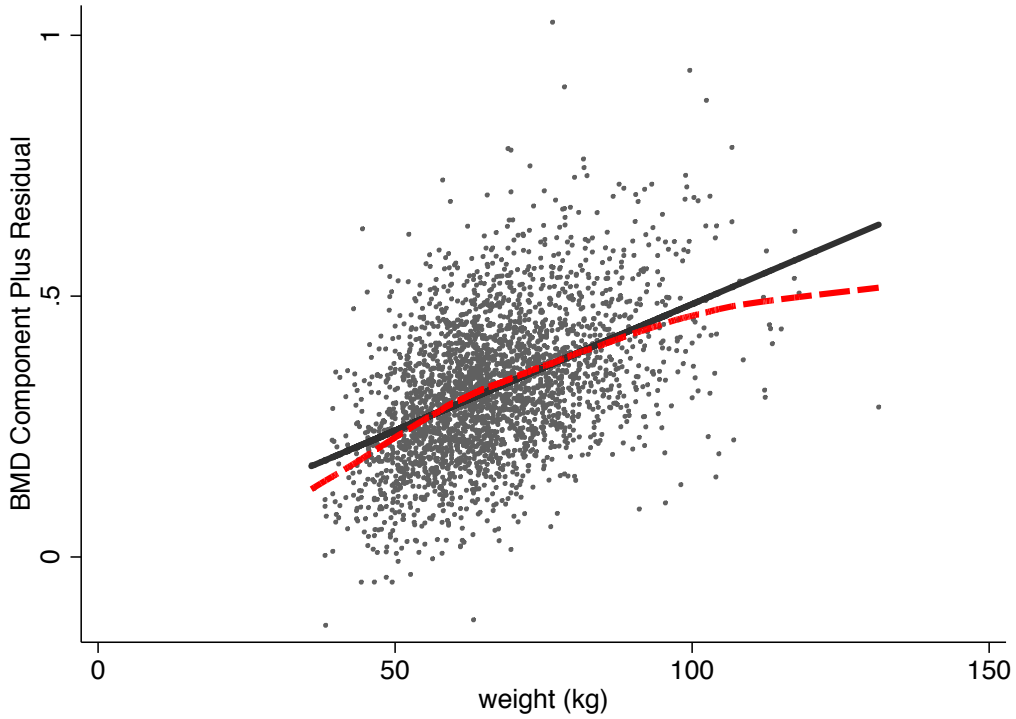
Diagnostics: RVP and CPR plots

- 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
  - e.g., $\mathrm{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
- In both CPR and RVP: mismatch of linear regression line, LOWESS smooth indicates lack of linearity

RVP plot for weight and BMD



CPR plot for weight and BMD



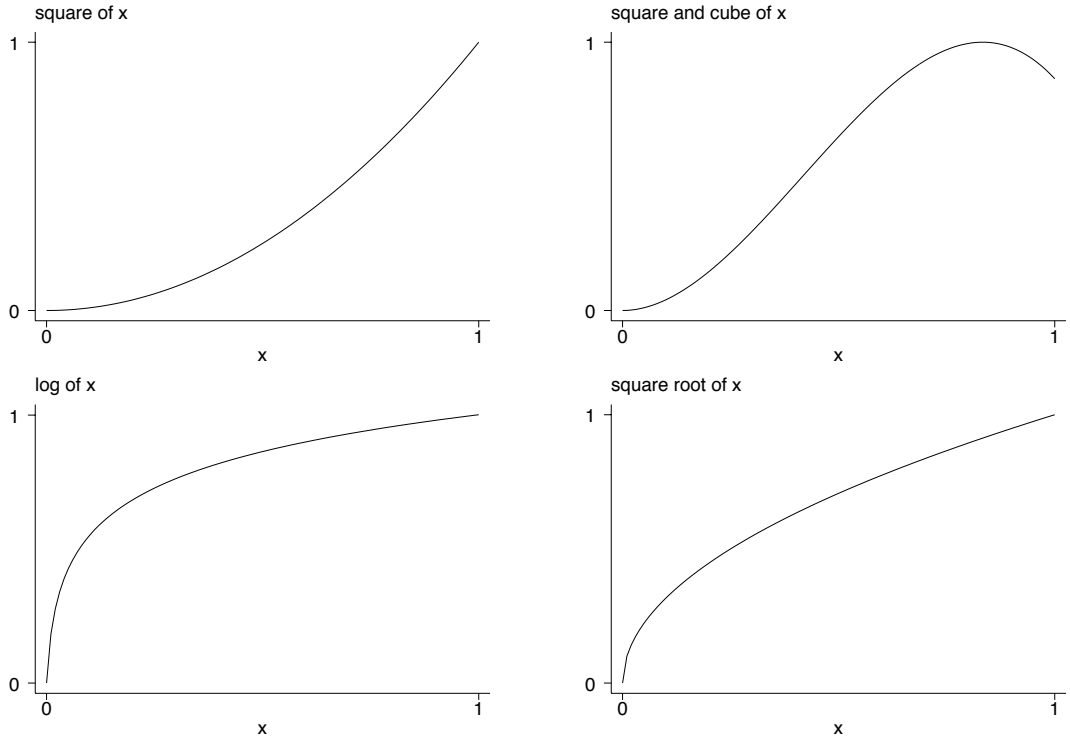
Stata code for RVP and CPR plots

```
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(s1mono) 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(s1mono) msize(tiny) ///
    ytitle("BMD Component Plus Residual") xtitle("weight (kg)") ///
    xlabel(0(50)150)
```

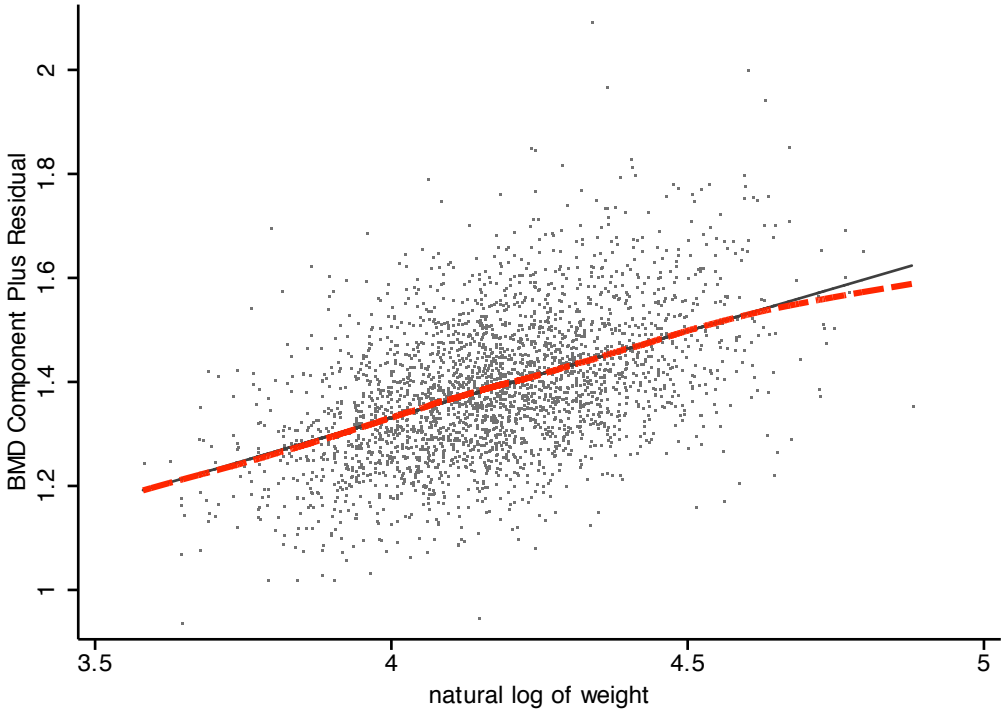
Solution: transform continuous predictors

- Smooth predictor transformations to fix non-linearity:
  - $\log(x)$ – provided $\mathrm{E}[Y|\mathbf{X}]$ is "monotone"
  - square root, cube root, other fractional powers of x
  - x^2, x^3 (lower order terms usually included in the model)

Predictor transformations



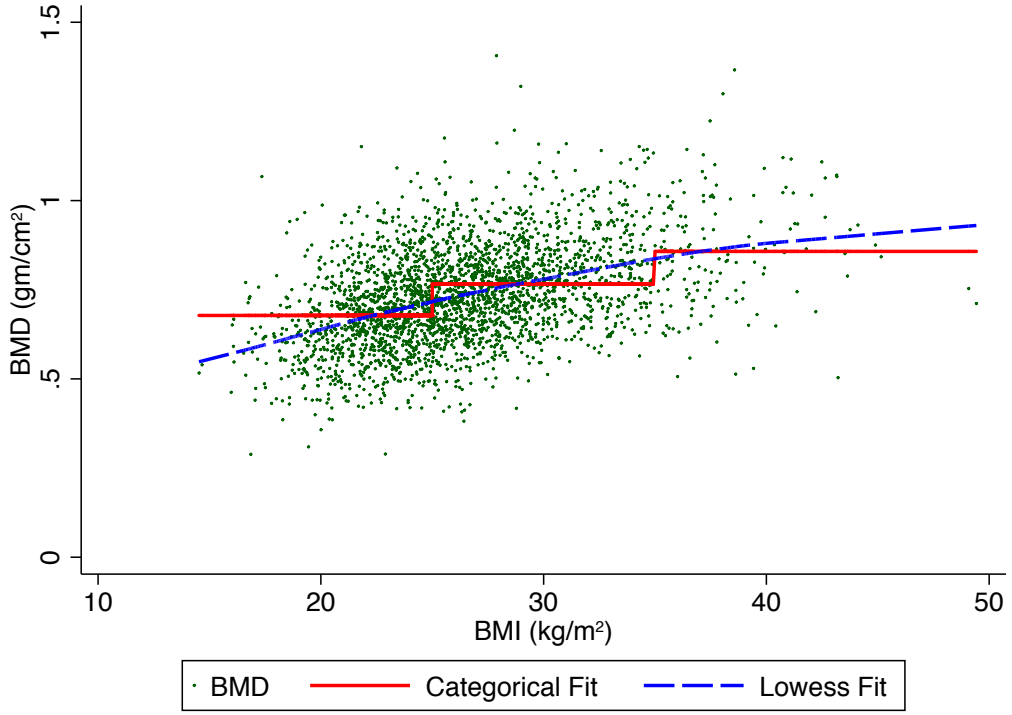
CPR plot for log-weight and BMD



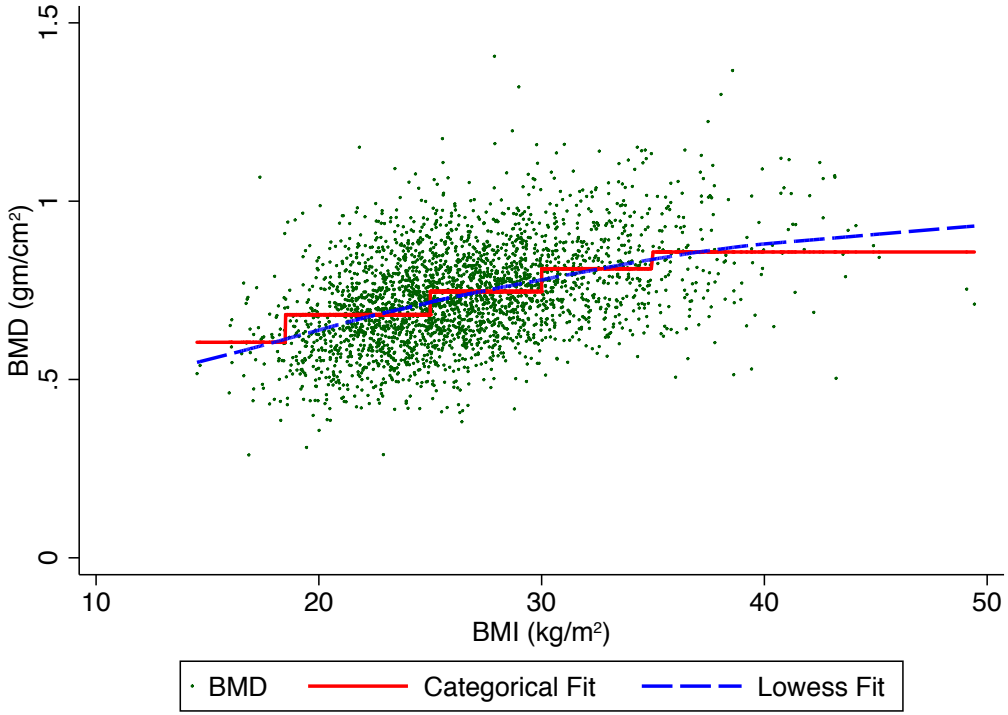
Alternatives: categorize the predictor

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

Too coarsely categorizing the predictor



A better tradeoff



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/cm{superscript:2})")  ///
    xtitle("BMI (kg/m{superscript:2})") ///
    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/cm{superscript:2})") ///
    xtitle("BMI (kg/m{superscript:2})") ///
    legend(label(2 "Categorical Fit") label(3 "Lowess Fit") rows(1))
```

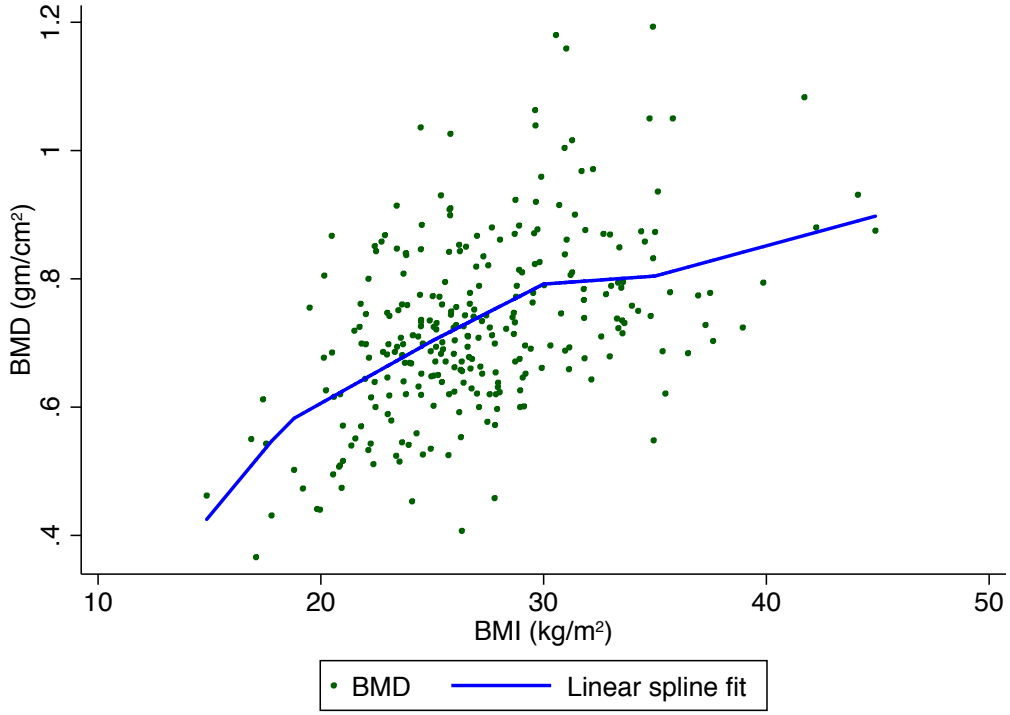
Alternatives: linear, restricted cubic splines

- *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)

Linear spline model for BMI effect on BMD

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

. regress bmd bmi1-bmi5

      Source |       SS       df       MS              Number of obs =     278
-------------+------------------------------           F(  5,   272) =   18.91
       Model |  1.34269169     5  .268538337           Prob > F      =  0.0000
    Residual |  3.86165215   272  .014197251           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
------------------------------------------------------------------------------
```

Linear spline fit



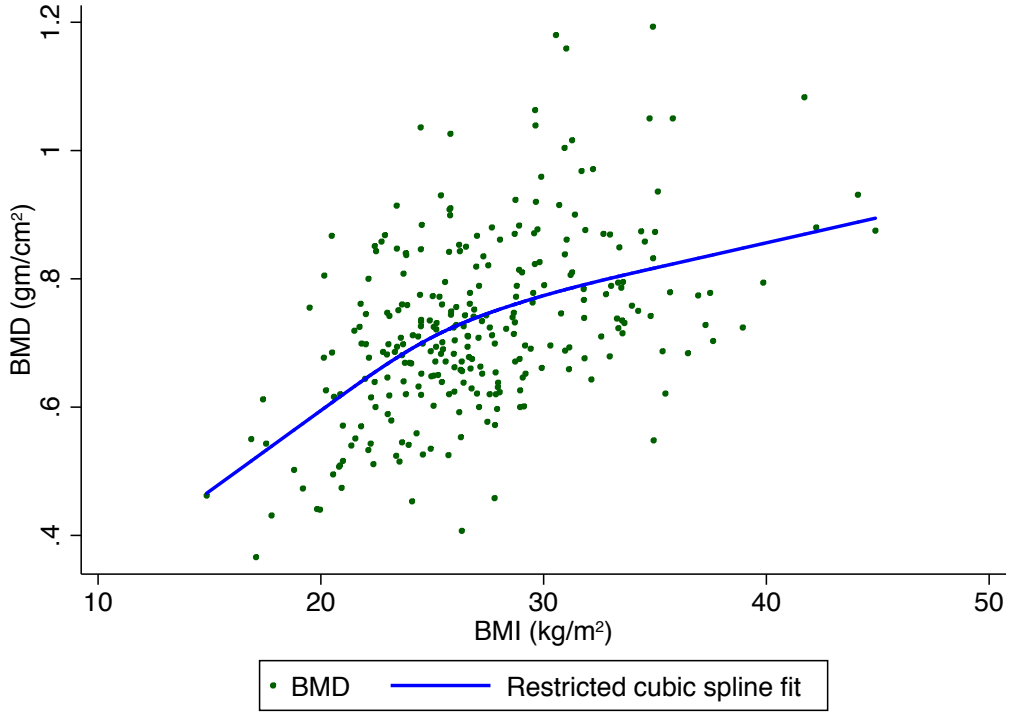
Tests for overall effect and non-linearity using linear spline model

```
. * test for overall effect
. testparm bmi*
 ( 1)  bmi1 = 0
 ( 2)  bmi2 = 0
 ( 3)  bmi3 = 0
 ( 4)  bmi4 = 0
 ( 5)  bmi5 = 0
       F(  5,   272) =   18.91
            Prob > F =    0.0000

. * test for departure from linearity
. 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
```

Cubic spline model for BMI effect on BMD

```
. mkspline bmircs = bmi, cubic nk(4) displayknots

             |     knot1      knot2      knot3      knot4
-------------+----------------------------------------
         bmi |  20.22235   24.99865   28.13695    35.4765

. regress bmd bmircs*

      Source |       SS       df       MS              Number of obs =     278
-------------+------------------------------           F(  3,   274) =   31.05
       Model |  1.32027072     3  .440090238           Prob > F      =  0.0000
    Residual |  3.88407312   274  .014175449           R-squared     =  0.2537
-------------+------------------------------           Adj R-squared =  0.2455
       Total |  5.20434383   277  .018788245           Root MSE      =  .11906

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bmircs1 |   .0251572   .0061574     4.09   0.000     .0130354    .0372789
     bmircs2 |  -.0244138   .0232578    -1.05   0.295    -.0702006     .021373
     bmircs3 |   .0488008   .0783588     0.62   0.534    -.1054611    .2030627
       _cons |   .0914822   .1377547     0.66   0.507    -.1797099    .3626743
------------------------------------------------------------------------------
. * Note: one fewer spline variables than number of knots
```

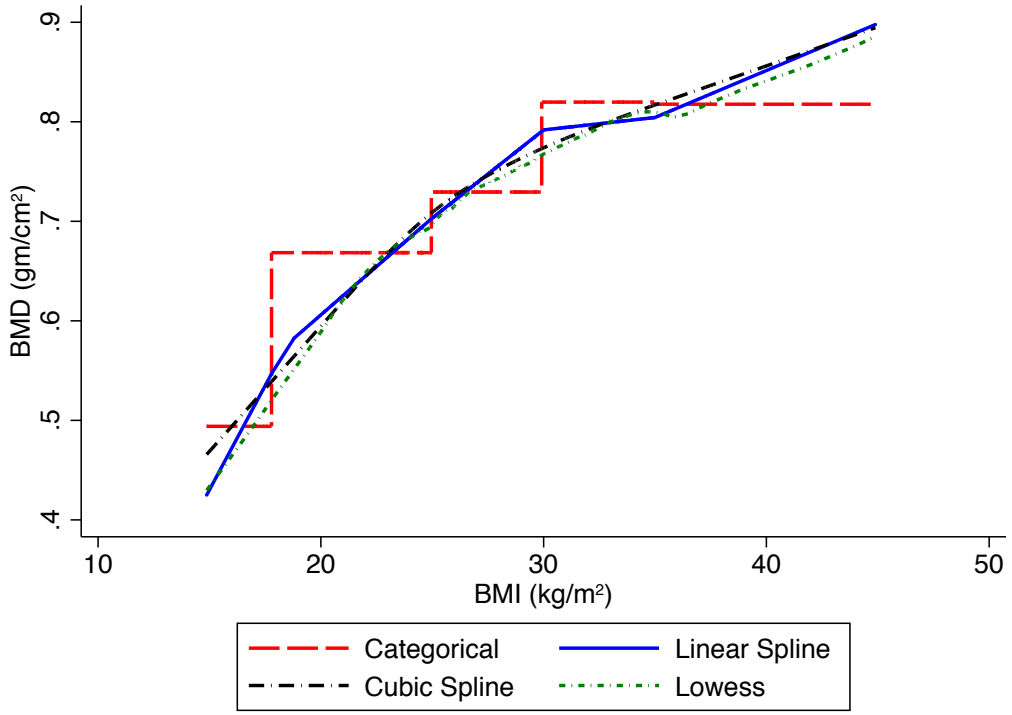
Cubic spline fit



Tests for overall effect, non-linearity

```
. * test for overall effect
. testparm bmircs*
 ( 1)  bmircs1 = 0
 ( 2)  bmircs2 = 0
 ( 3)  bmircs3 = 0
       F(  3,   274) =   31.05
            Prob > F =    0.0000

. * test for departure from linearity
. testparm bmircs2 bmircs3
 ( 1)  bmircs2 = 0
 ( 2)  bmircs3 = 0
       F(  2,   274) =    3.69
            Prob > F =    0.0262
```

Categorical, linear and cubic spline fits



Stata code for plot

```
twoway ///
       (line fitted bmi, sort lp(longdash) lcol(red) lw(medthick) co(stepstair)) ///
       (line fitted2 bmi, sort lp(solid) lcol(blue) lw(medthick)) ///
       (line fitted3 bmi, sort lp(dash_dot) lcol(black) lw(medthick)) ///
       (lowess bmd bmi, lp(shortdash_dot) lcol(green) lw(medthick)), ///
       plotregion(style(none)) scheme(s1color) ///
       xtitle("BMI (kg/m{superscript:2})") ///
       ytitle("BMD (gm/cm{superscript:2})") ///
       legend(label(1 "Categorical") label(2 "Linear Spline") ///
              label(3 "Cubic Spline") label(4 "Lowess") rows(2))
```

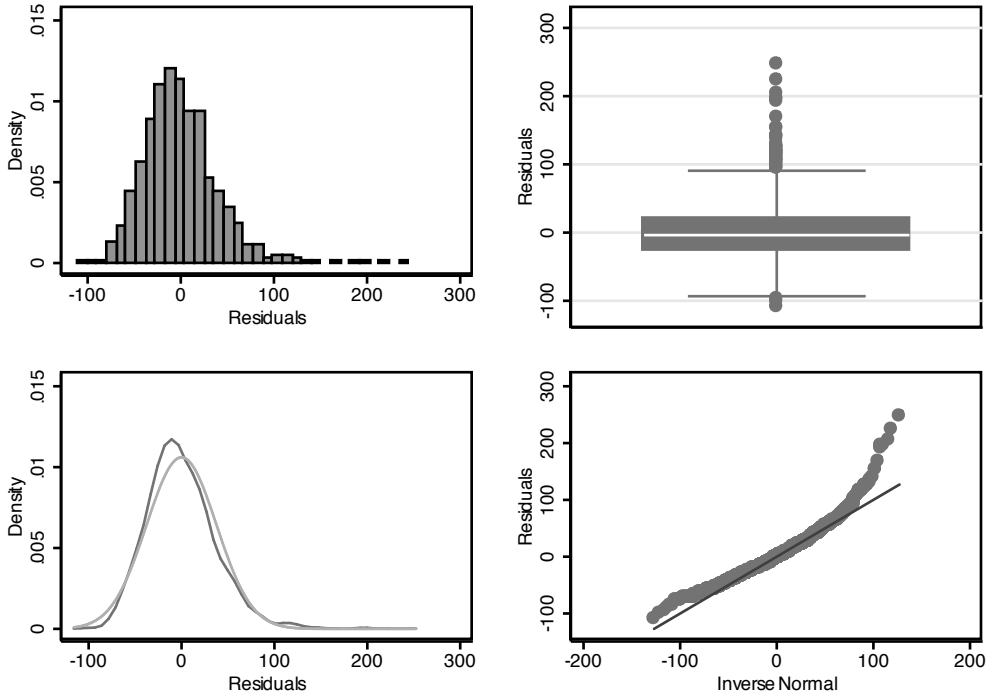
Checking linearity: summary

- Diagnostics:
  - linear models: curved LOWESS smooth in CPR or RVP plot
  - linear, logistic, Cox models: fit restricted cubic spline, test for departure from linearity using `testparm` for all but first spline component
- Solutions: transform predictor, possibly using linear or cubic splines
- Note: regression models *do not require predictor normality*

Checking the model: normality

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

Diagnosing departures from normality



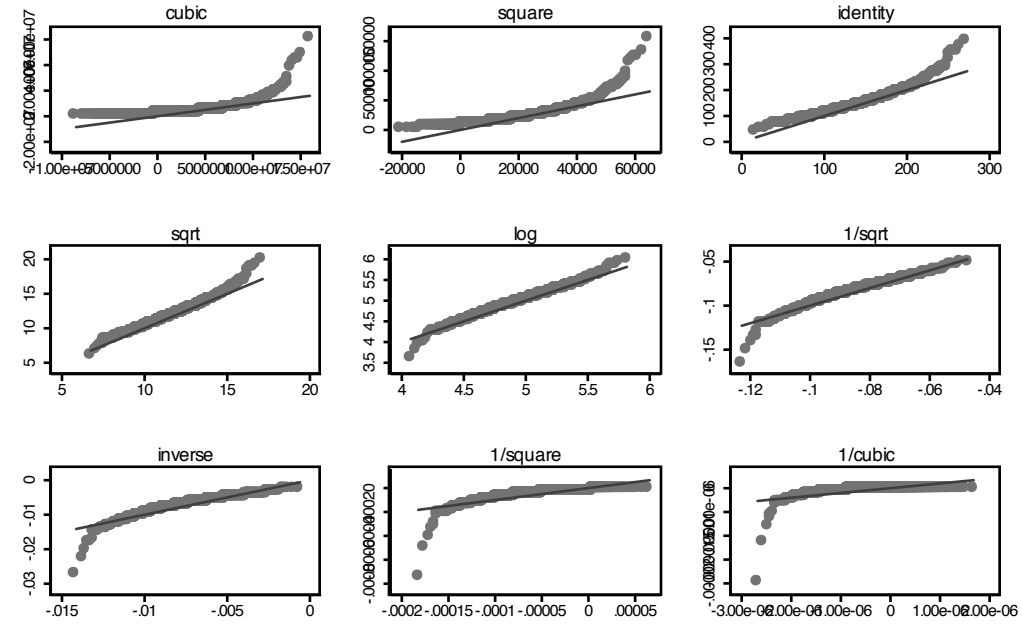
Stata code for normality checks

```
program define eda
set graphics off
set scheme s1mono
quietly histogram ‘1’, name(eda1, replace)
quietly graph box ‘1’, name(eda2, replace)
quietly kdensity ‘1’, ep normal legend(off) name(eda3, replace)
quietly qnorm ‘1’, name(eda4, replace)
set graphics on
set scheme s1mono
graph combine eda1 eda2 eda3 eda4, name(eda, replace)
end

regress ldl bmi age nonwhite smoking drinkany
predict resid, resid
eda resid
```

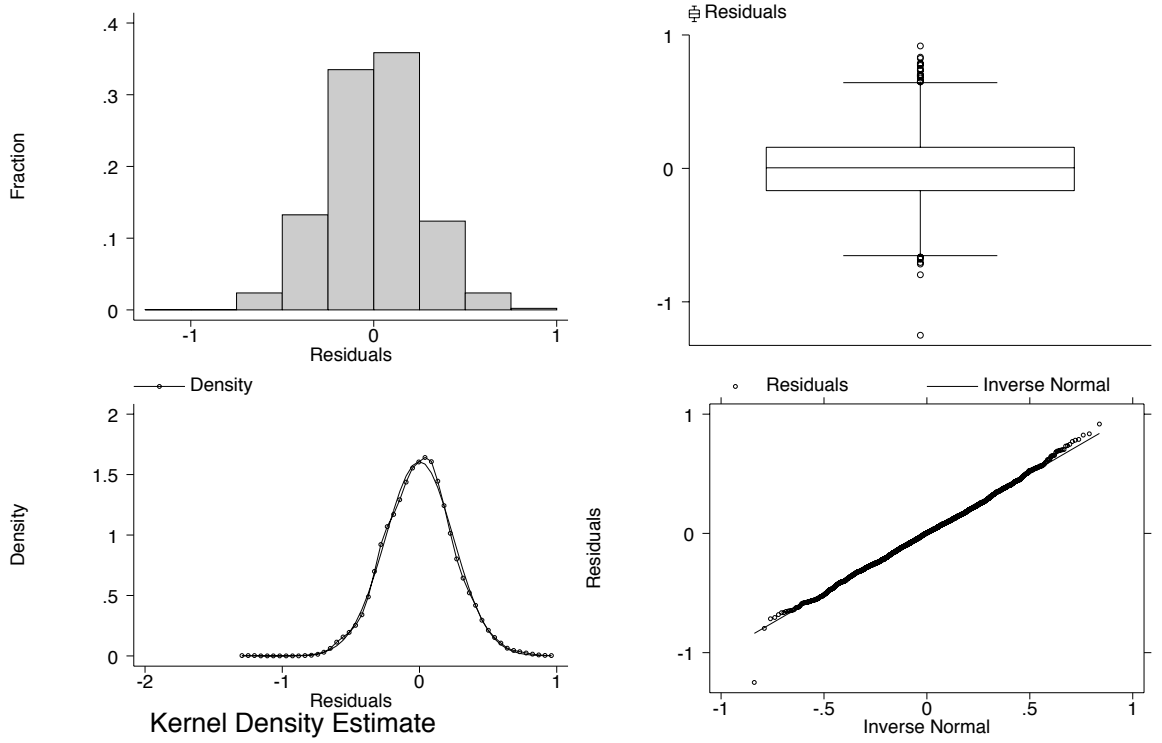
Solution: transform the outcome

- Residuals skewed (usually to the right):
  - log, square root, other power transformations
  - may need to add constant to make all values positive
- Search for best transformation using `gladder` command
  - note: `gladder` applied to outcome, not residuals
- Residuals symmetrically long-tailed
  - rank transformation, trimming, Winsorization

G-ladder plots for LDL



Residuals of log-transformed LDL



Another solution: bootstrap CIs

- Resample N observations *with replacement* from data, re-fit model, store estimates, repeat

- Distribution of bootstrap estimates models sampling distribution of actual estimate

- Quick, partial solution:

  1. replace model-based SE by SD of bootstrap estimates

  2. construct CIs assuming Normality

A better solution: percentile bootstrap CIs

- 95% CI: 2.5^{th} to 97.5^{th} percentile of bootstrap estimates
- Bias-correction shifts CI slightly to right or left
- Slower but avoids making Normality assumption
- Requires using many (≥ 500) bootstrap samples
  - extreme percentiles are noisy!
  - default is 50

Stata code for normal based bootstrap CIs

```
. regress LDL BMI age nonwhite smoking drinkany, vce(bootstrap)

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

Linear regression                               Number of obs      =      2745
                                                Replications       =        50
                                                Wald chi2(5)       =     31.14
                                                Prob > chi2        =    0.0000
                                                R-squared          =    0.0108
                                                Adj R-squared      =    0.0090
                                                Root MSE           =   37.6467
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
         LDL |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .3591038    .147644     2.43   0.015      .069727    .6484807
         age |  -.1897166   .1048131    -1.81   0.070    -.3951465    .0157133
    nonwhite |   5.219436   2.379216     2.19   0.028     .5562581    9.882614
     smoking |   4.750738   2.625508     1.81   0.070    -.3951626    9.896639
    drinkany |  -2.722354   1.336584    -2.04   0.042     -5.34201   -.1026974
       _cons |   147.3153   8.531092    17.27   0.000     130.5947    164.0359
------------------------------------------------------------------------------
```

Stata code for bias-corrected percentile CIs

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

. estat bootstrap

Linear regression                               Number of obs      =      2745
                                                Replications       =      1000

------------------------------------------------------------------------------
             |    Observed                 Bootstrap
         LDL |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .35910384   .0037356   .12982959    .1288215   .6207935  (BC)
         age |   -.1897166   .0037144   .11026674   -.4109758   .0241487  (BC)
    nonwhite |   5.2194359  -.0446811   2.5560214     .707088   10.63076  (BC)
     smoking |    4.750738   .0401205    2.203131    .3956041   9.098602  (BC)
    drinkany |  -2.7223535    .060899   1.5013767   -5.880945  -.1049485  (BC)
       _cons |   147.31529  -.4018183   9.0308581    130.2375   165.5541  (BC)
------------------------------------------------------------------------------
(BC)   bias-corrected confidence interval
```

Solution: Generalized Linear Models (GLMs)

- GLMs model a transform of the mean (rather than a transform of the outcome), do not require normality

- Logistic model for binary outcomes uses *logit* transformation of the mean $\mathrm{E}[Y|\mathbf{X}] = Pr[Y = 1|\mathbf{X}]$

$$\log \frac{\mathrm{E}[Y|\mathbf{X}])}{1 - \mathrm{E}[Y|\mathbf{X}]} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Other generalized linear models (GLMs) avoid dichotomizing outcome, generally model $\log \mathrm{E}[Y|\mathbf{X}]$ (Biostat 209)

  – gamma, Poisson, negative binomial, zero-inflated models

Solution: ordinal models

- 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 models (continued)

- Alternatively, categorize CAC as 0, 1-9, 10-99, 100-399, $\geq$ 400, use regression model for ordinal outcomes
  - proportional odds (`ologit`)
  - continuation ratio (`ocratio`)
- Proportional odds assumption relaxed using `gologit2`
- Steve will briefly cover these

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

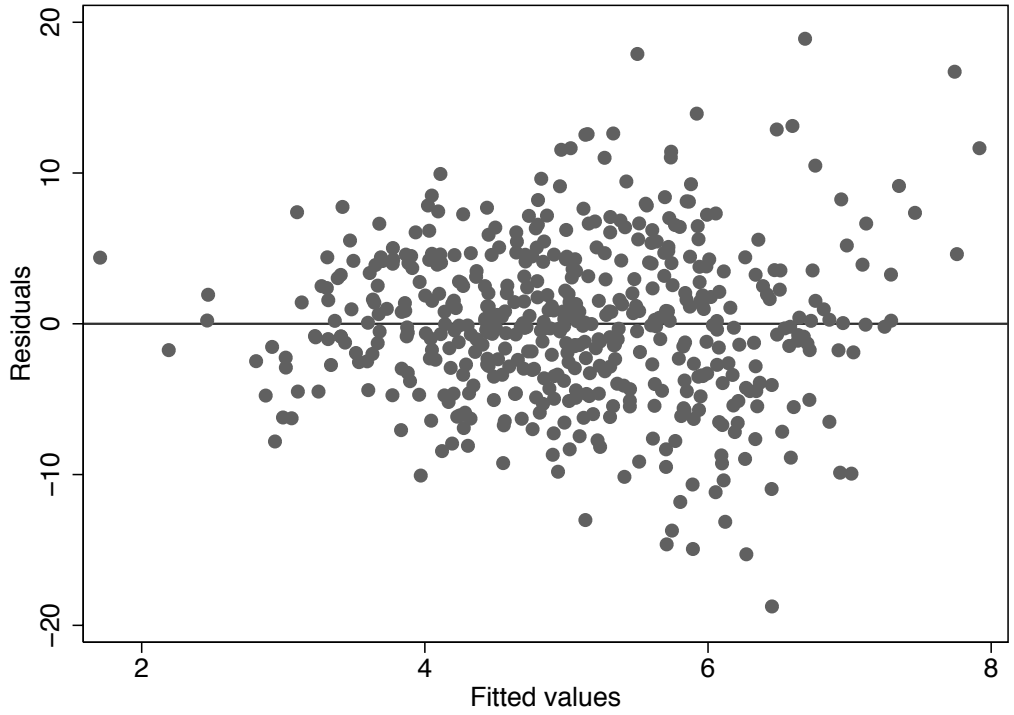
Checking the model: constant variance

- If constant variance assumption is violated
  - coefficient estimates unbiased but inefficient
  - tests for between-group differences may be invalid
  - unlike Normality problems, larger samples don't help

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

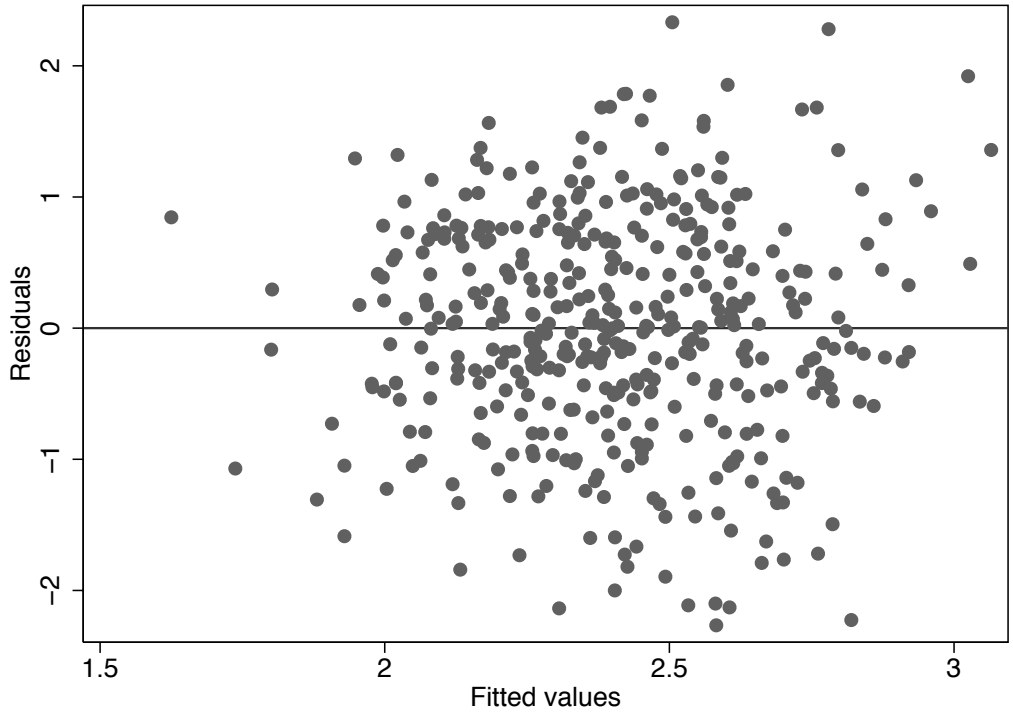
RVF plot to diagnose non-constant variance



Solution: transform outcome

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

After square root transformation of outcome



Comparing N, residual variance by subgroup

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

         physact |         N  variance
-----------------+--------------------
much less active |        26  1198.729
somewhat less ac |        46  746.4037
 about as active |        87  990.6615
somewhat more ac |        85   527.047
much more active |        32  124.3417
-------------------------------------

. tabstat resid, by(diabetes) stat(n var) nototal

diabetes |         N  variance
---------+--------------------
      no |       196   100.288
     yes |        80  2244.603
-----------------------------
```

Solution: use robust SEs

```
. regress glucose diabetes i.physact age i.raceth smoking drinkany, vce(robust)
......
------------------------------------------------------------------------------
             |               Robust
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    diabetes |   55.32816   5.704065     9.70   0.000     44.09711    66.55922
             |
     physact |
           2 |   .5986391   7.670311     0.08   0.938    -14.50387    15.70115
           3 |    6.51184   7.519767     0.87   0.387    -8.294252    21.31793
           4 |   2.873804   7.282648     0.39   0.693    -11.46541    17.21302
           5 |   .4625191   6.907942     0.07   0.947    -13.13892    14.06396
             |
         age |  -.3130465   .2466262    -1.27   0.205    -.7986428    .1725497
             |
      raceth |
           2 |   9.907849   7.805314     1.27   0.205    -5.460473    25.27617
           3 |   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
------------------------------------------------------------------------------
```

... or use more conservative robust SEs

```
. regress glucose diabetes i.physact age i.raceth smoking drinkany, vce(hc3)
 .....
------------------------------------------------------------------------------
             |             Robust HC3
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    diabetes |   55.32816   5.838609     9.48   0.000      43.8322    66.82413
             |
     physact |
           2 |   .5986391   8.082405     0.07   0.941    -15.31526    16.51254
           3 |    6.51184   7.877636     0.83   0.409    -8.998881    22.02256
           4 |   2.873804   7.619965     0.38   0.706    -12.12957    17.87718
           5 |   .4625191   7.247594     0.06   0.949    -13.80768    14.73271
             |
         age |  -.3130465   .2557038    -1.22   0.222    -.8165162    .1904231
             |
      raceth |
           2 |   9.907849   8.189902     1.21   0.227     -6.21771    26.03341
           3 |   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
------------------------------------------------------------------------------
```

Solution: use GLMs with robust standard errors

Distribution	Variance-to-Mean Relationship	Outcome
Normal	σ^2 constant	Continuous
Binomial	$\sigma^2 = n\mu(1-\mu)$	Successes in n trials
OD* Binomial	$\sigma^2 \propto n\mu(1-\mu)$	Clustered successes
Poisson	$\sigma^2 = \mu$	Counts
OD* Poisson	$\sigma^2 \propto \mu$	Counts
Negative binomial	$\sigma^2 = \mu + \mu^2/k$	Counts
Gamma	$\sigma \propto \mu$	Continuous

* over-dispersed

See Table 8.8, VGSM; covered in Biostat 209

Checking constant variance: summary

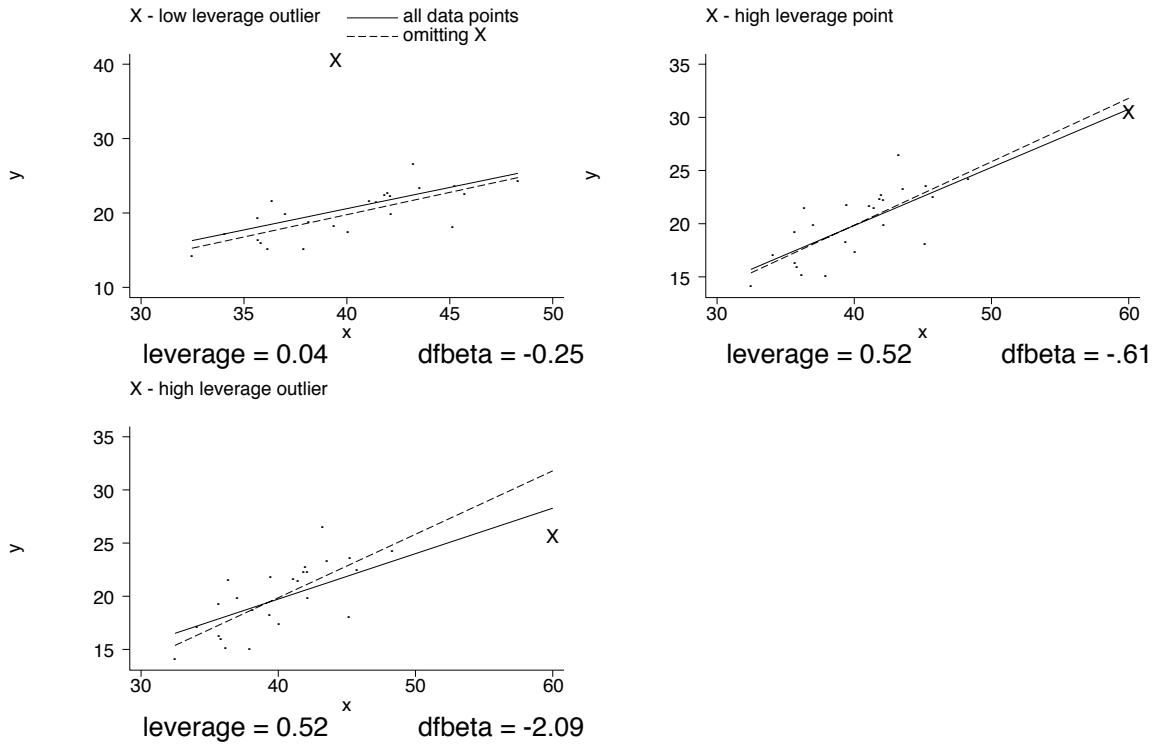
- Diagnostics: funnel shapes in RVP plot, variable Ns, SDs across subgroups

- Solutions: transform outcome, use robust SEs or GLM

Checking the model: high leverage and influential points

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

Simple outlier, high leverage, high influence



Diagnostics: boxplots of `dfbeta` statistics

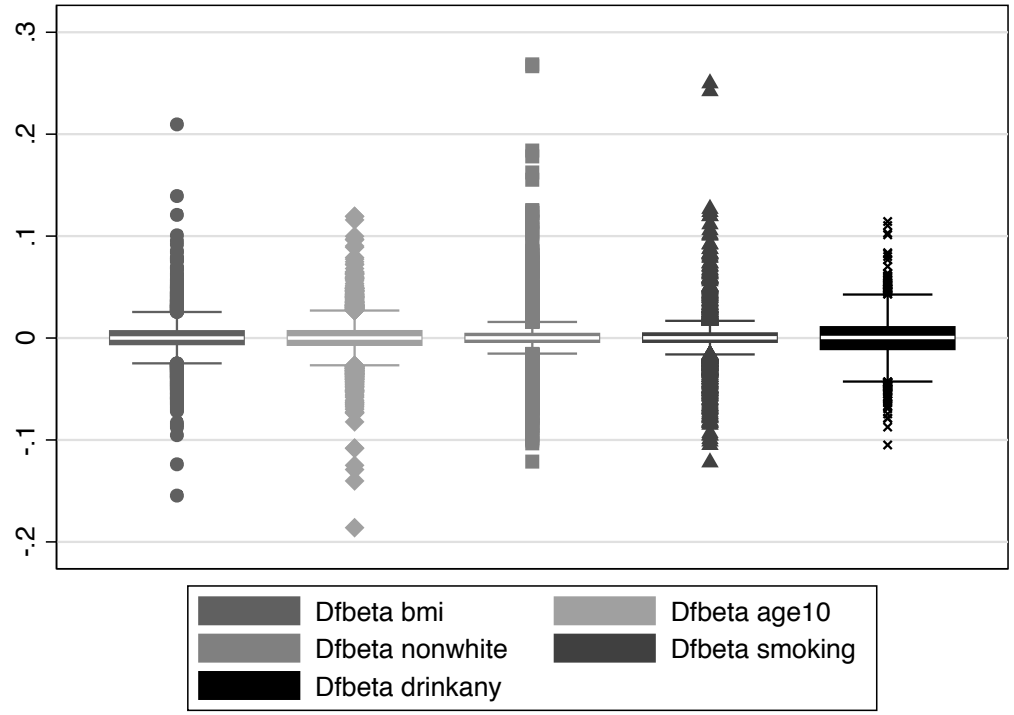
- `dfbeta` statistics measure changes in each β_j when each data point is omitted

- Defined for each observation and predictor in model

- Check for outliers in boxplots of `dfbeta`s

```
reg ldl bmi age10 nonwhite smoking drinkany
dfbeta
graph box _dfbeta_*
```

Boxplots of dfbetas for BMI - LDL model



Solution

- Identify up to 10 observations with biggest DFbetas
- Check for data errors or other anomaly
- Refit model without influential points, re-assess conclusions, report sensitivities
- Consider deleting influential points if they represent a different population
- Note: suggested rules for defining influential points in Stata documentation are too strict

Sensitivity of LDL model to 4 influential points with `dfbetas`$>$0.2 in absolute value

Predictor variable	All observations		Omitting 4 points	
	$\widehat{\beta}$	P-Value	$\widehat{\beta}$	P-Value
BMI	0.36	0.007	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

Checking the model: covariate overlap

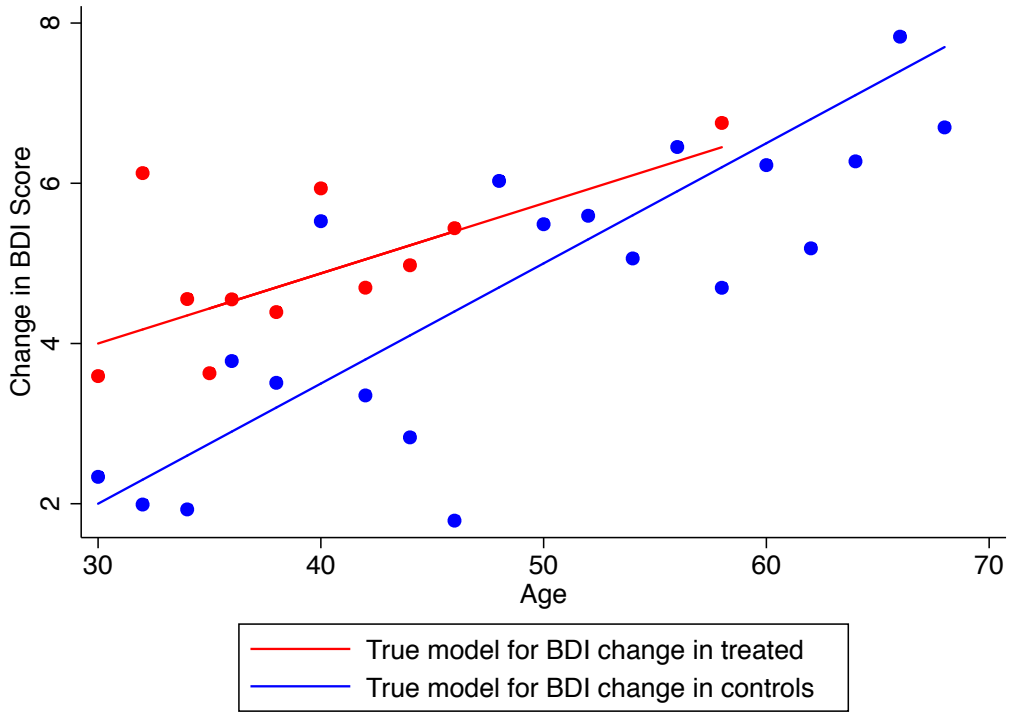
- Observational analysis of effect of binary exposure problematic if exposed, unexposed too unlike

- Lack of overlap makes true model hard to find, especially in small datasets

- Comparing each covariate in exposed and unexposed may not be enough, because covariates are correlated:

  - some *combinations* of covariates may be unrepresented in one group

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



No power to detect interaction

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

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

Diagnosing lack of overlap

- Compare distribution of covariates in exposed and unexposed

- Use propensity scores
  - fit logistic model for binary primary predictor
    - adjust for minimum sufficient adjustment set (MSAS) for the exposure-outcome relationship, capturing non-linearities and interactions
  - get fitted values (on linear predictor or probability scale)
  - plot the results by primary predictor and check overlap

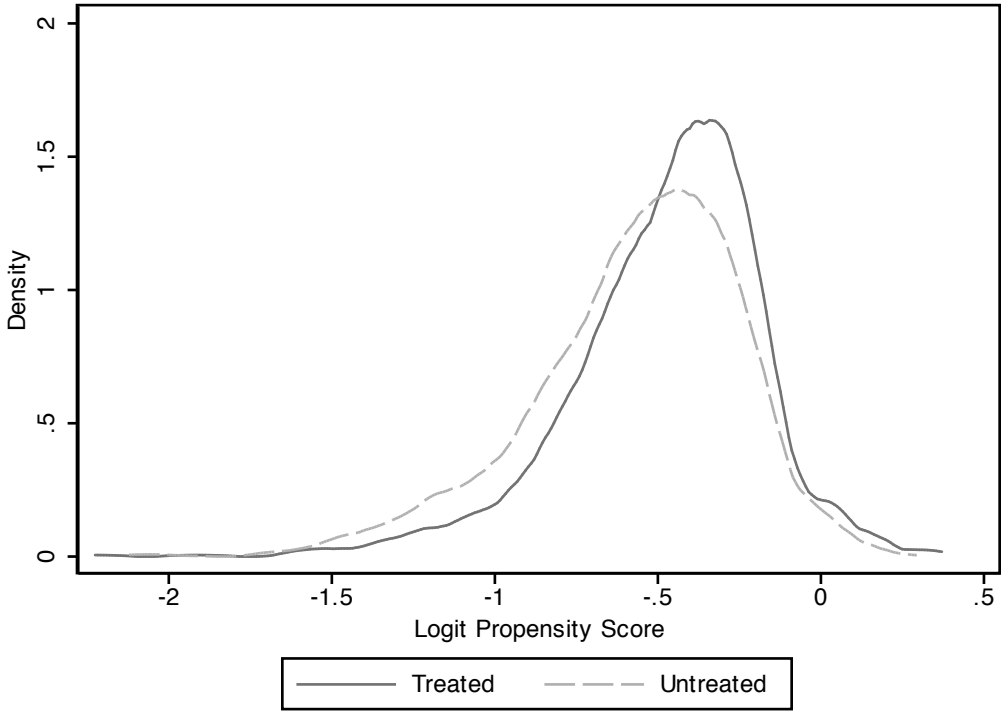
Comparing covariate distributions by statin use

```
. tabstat age, by(statins) s(n mean sd min max) format(%5.2f)
statins |         N      mean        sd       min       max
--------+----------------------------------------------------
     no |   1732.00     66.53      6.90     44.00     79.00
    yes |   1031.00     66.85      6.22     45.00     79.00
--------+----------------------------------------------------
  Total |   2763.00     66.65      6.65     44.00     79.00
----------------------------------------------------------------

. tab statins educ_cat, row nokey

    use of |         RECODE of educyrs (years of education)
   statins |       <HS    HS grad        >HS         BS        >BS |     Total
-----------+---------------------------------------------------------+----------
        no |       380        676        416        136        123 |     1,731
           |     21.95      39.05      24.03       7.86       7.11 |    100.00
-----------+---------------------------------------------------------+----------
       yes |       185        415        252         77        101 |     1,030
           |     17.96      40.29      24.47       7.48       9.81 |    100.00
-----------+---------------------------------------------------------+----------
     Total |       565      1,091        668        213        224 |     2,761
           |     20.46      39.51      24.19       7.71       8.11 |    100.00
```

Propensity score model for statin use

```
* 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"))
```

Overlap diagnostics for statin use

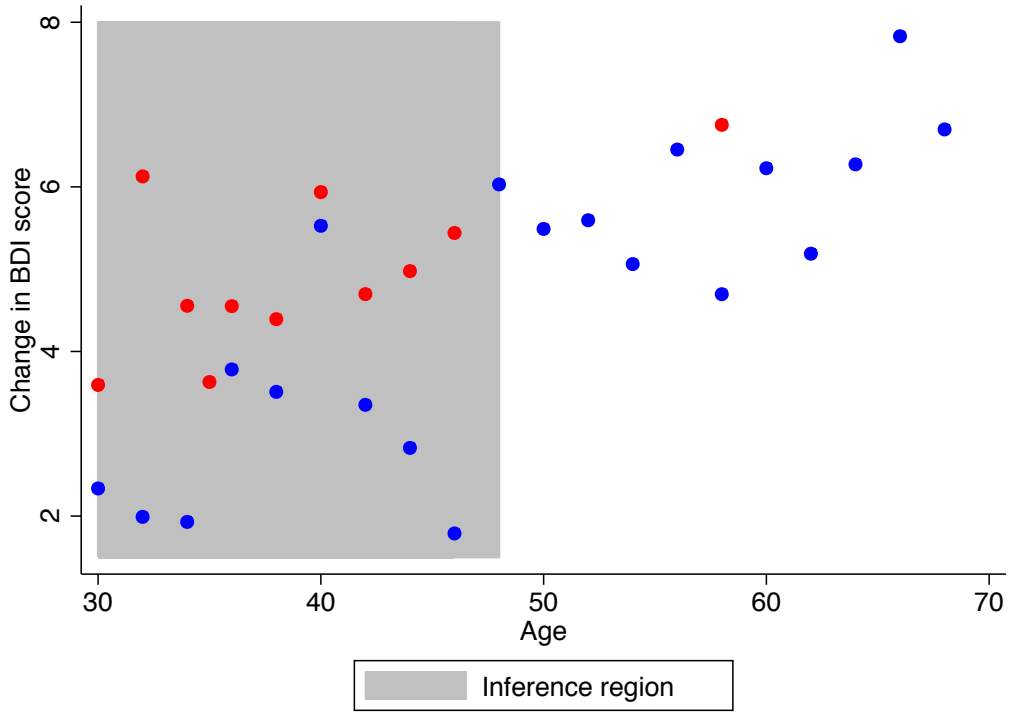


Solution: lack of overlap

- Restrict inference to region of good overlap

- Match on prognostic covariates or propensity scores

Restricting inference to region of overlap



Checking overlap: summary

- Diagnostics: compare covariates, density plots of logit-propensity scores in exposed, unexposed

- Solutions: restrict inference to region of good overlap, possibly by matching

Model checking: to 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-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

- 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 `dfbeta` statistics
  - Solutions: identify up to 10 influential points, correct data errors, omit influential points if justifiable, present sensitivity analysis