Generalizing diagnostics from single-variance linear models

Many papers have proposed individual methods to look for outliers, check normality of residuals, detect influential observations, etc.

I know of two attempts to produce a system of diagnostics for MLMs.

One is Yuan & Johnson (2012 *Biometrics*).

I'll present the other one – Hodges (1998 *JRSSB*).

RWC and Lee et al (LNP) use residuals to assess non-linearity and non-constant error variance.

RWC seem unaware of problems with residuals (discussed below); LNP acknowledge them but blow them off.

I'll focus on Hodges (1998) with brief comments on RWC and LNP.

Some preliminary items

The main analysis is Bayesian; assume we have MCMC draws.

I follow Weisberg (1983) in seeking these qualities in diagnostics:

- A diagnostic should aim to detect a specific problem.
- Diagnostics should compute quickly.
- A diagnostic should have a corresponding plot, so you can assess the effect of individual observations.
- Graphics should let users to look at the data as directly as possible.

The last item argues against using a Bayesian approach to *diagnostics*, even when using a Bayesian approach to *analysis*.

This lecture uses the constraint-case formulation.

We'll use "the HMO dataset"

Assembled in the mid-1990s, it describes 341 HMOs serving U.S. Government employees.

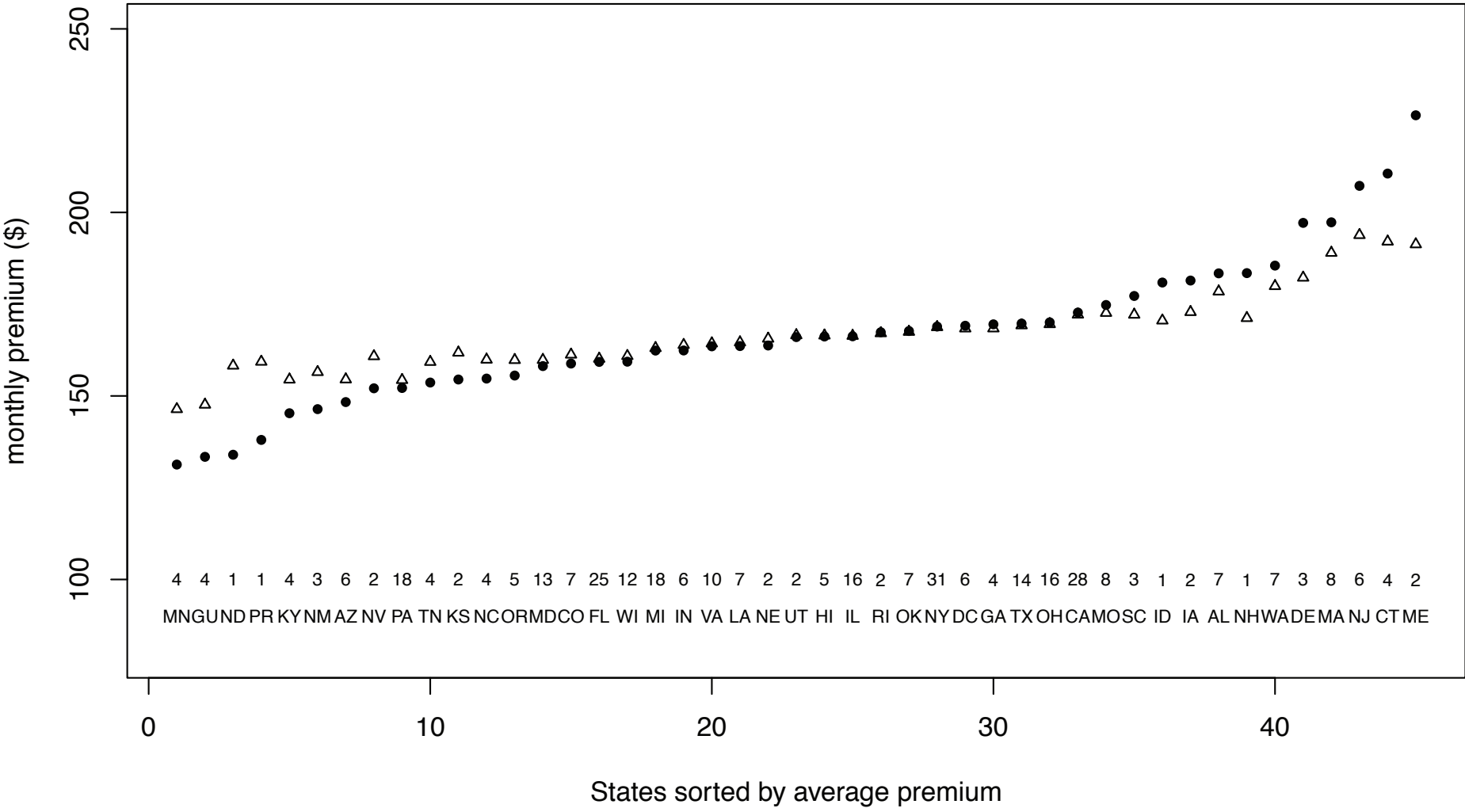
It has plans in 42 states, DC, Guam, and Puerto Rico, all called "states."

- Between 1 and 31 plans per state, median 5.

Of particular interest: each HMO's monthly premium ($) for individuals.

- We have some plan-level and state-level covariates.

The next slide summarizes the data.



For each state: the numeral is its number of plans, • is its average premium, △ is its posterior mean in the simple analysis (next).

Begin with an unbalanced one-way RE model:

$$\begin{aligned} y_{ij} &= \theta_i + \epsilon_{ij}, \quad \text{for state } i = 1, \ldots, 45 \\ \theta_i &= \mu + \delta_i, \quad \text{for plan within state } j = 1, \ldots, n_i \end{aligned} \tag{1}$$

$$\epsilon_{ij} \sim N(0, \sigma_e^2),\ \delta_i \sim N(0, \sigma_s^2);\ \text{flat prior on } \mu.$$

Rewrite (1) as $0 = -\theta_i + \mu + \delta_i$ to give the constraint-case form:

$$\left[\begin{array}{c} \mathbf{y} \\ \hline \mathbf{0}_{45\times 1} \end{array}\right] = \left[\begin{array}{ccc|c} \mathbf{1}_{n_1} & \cdots & \mathbf{0}_{n_1} & \\ \vdots & \ddots & \vdots & \mathbf{0}_{341\times 1} \\ \mathbf{0}_{n_{45}} & \cdots & \mathbf{1}_{n_{45}} & \\ \hline \multicolumn{3}{c|}{-\mathbf{I}_{45}} & \mathbf{1}_{45} \end{array}\right] \left[\begin{array}{c} \theta_1 \\ \vdots \\ \theta_{45} \\ \hline \mu \end{array}\right] + \left[\begin{array}{c} \epsilon \\ \hline \delta \end{array}\right]$$

$$\text{cov}\left[\begin{array}{c} \epsilon \\ \hline \delta \end{array}\right] = \left[\begin{array}{c|c} \sigma_e^2 \mathbf{I}_{341} & \mathbf{0} \\ \hline \mathbf{0} & \sigma_s^2 \mathbf{I}_{45} \end{array}\right]$$

$\mathbf{y}$ and ϵ are 341×1, ordered with j changing fastest; δ is 45×1.

Here's the simple model for the HMO data again:

$$\left[\begin{array}{c}\mathbf{y}\\\hline \mathbf{0}_{45\times 1}\end{array}\right] = \left[\begin{array}{ccc|c} \mathbf{1}_{n_1} & \cdots & \mathbf{0}_{n_1} & \\ \vdots & \ddots & \vdots & \mathbf{0}_{341\times 1} \\ \mathbf{0}_{n_{45}} & \cdots & \mathbf{1}_{n_{45}} & \\ \hline \multicolumn{3}{c|}{-\mathbf{I}_{45}} & \mathbf{1}_{45} \end{array}\right] \left[\begin{array}{c}\theta_1\\ \vdots \\ \theta_{45}\\\hline \mu\end{array}\right] + \left[\begin{array}{c}\epsilon\\\hline \delta\end{array}\right]$$

To derive and present results in more generality, I'll use the form

$$\left[\begin{array}{c}\mathbf{y}\\\hline \mathbf{0}\\\hline \mathbf{M}\end{array}\right] = \left[\begin{array}{c|c}\mathbf{K} & \mathbf{0}\\\hline \mathbf{S}_1 & \mathbf{S}_2\\\hline \mathbf{W}_1 & \mathbf{W}_2\end{array}\right] \left[\begin{array}{c}\Theta_1\\\hline \Theta_2\end{array}\right] + \left[\begin{array}{c}\epsilon\\\hline \delta\\\hline \xi\end{array}\right],$$

I'll summarize this as $\mathbf{Y} = \mathbf{H}\Theta + \mathbf{E}$, where $\text{cov}(\mathbf{E})$ has a simple form.

Added variable plots (AVPs) in the usual linear model

An AVP displays the evidence for adding a predictor to a linear model.

Suppose we've included predictors in $\mathbf{A}$ and are considering $\mathbf{B}$.

$$\text{The usual linear model is } \mathbf{y} = \mathbf{A}\beta + \mathbf{B}\omega + \epsilon.$$

Premultiply both sides of this by $\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$:

$$\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{B}\omega + (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\epsilon$$

If $E(\epsilon) = 0$ as usual, then

$$E(\hat{\mathbf{e}}) = (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{B}\omega.$$

The AVP plots $\hat{\mathbf{e}}$ vs $(\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{B}$.

If $\mathbf{B}$ should be added, this will fit a line through the origin with slope ω.

AVPs for mixed linear models in constraint-case form

Predictors can be added for data or constraint cases.

The AVP for linear models can be applied directly to either kind of predictor using four steps (explained on following slides)

1. Reformulate the candidate predictor and call it $\mathbf{B}$.
2. Write the model with the candidate predictor as $\mathbf{Y} = \mathbf{H}\Theta + \mathbf{B}\phi + \mathbf{E}$.
3. Pre-multiply that equation by $\mathbf{\Gamma}^{-1/2}$ for a suitable $\mathbf{\Gamma}$, to make it a homoscedastic-errors problem.
4. Draw the usual AVP.

Step 1. Reformulate the candidate predictor and call it $\mathbf{B}$.

If $\mathbf{B}_1$ is the 341-vector of plan enrollments, $\mathbf{B}$ is $[\mathbf{B}_1', \mathbf{0}_{1\times 45}]'$.

If $\mathbf{B}_2$ is the 45-vector of state populations, $\mathbf{B}$ is $[\mathbf{0}_{1\times 341}, \mathbf{B}_2']'$.

Step 3. Make it homoskedastic: The AVP is not too sensitive to $\mathbf{\Gamma}$.

Start with $\mathbf{Y} = \mathbf{H}\Theta + \mathbf{B}\omega + \mathbf{E}$; pre-multiply by $\mathbf{\Gamma}^{-1/2}$ to give

$$\mathsf{Y} = \mathsf{H}\Theta + \mathsf{B}\omega + \mathsf{E},$$

where sans-serif font indicates pre-multiplication by $\mathbf{\Gamma}^{-1/2}$.

Pre-multiply by $\mathbf{I} - \mathsf{H}(\mathsf{H}'\mathsf{H})^{-1}\mathsf{H}'$, to give

$$\hat{\mathsf{E}} = (\mathbf{I} - \mathbf{V})\mathsf{Y} = (\mathbf{I} - \mathbf{V})\mathsf{B}\omega + (\mathbf{I} - \mathbf{V})\mathsf{E}.$$

For any $\mathbf{\Gamma}$, the AVP will show the correct slope ω.

Step 4. Draw the usual added-variable plot.

This plot includes the data- and constraint-case residuals (and prior-case residuals, if present).

The AVP's vertical axis is

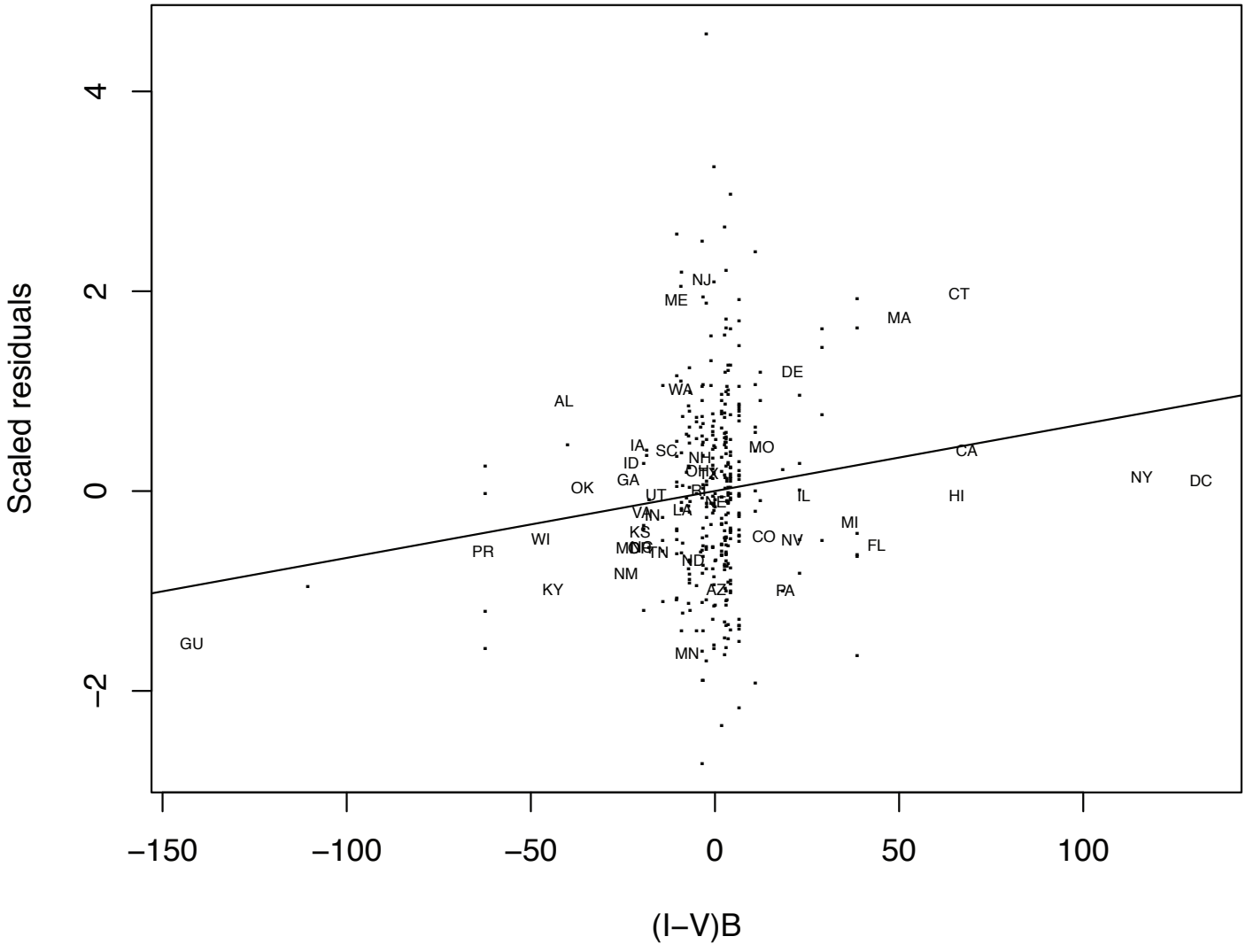
$$\hat{\mathsf{E}} = \boldsymbol{\Gamma}^{-1/2}\hat{\mathbf{E}} = \boldsymbol{\Gamma}^{-1/2}(\mathbf{Y} - \mathbf{H}\hat{\Theta}).$$

This scaling puts the different kinds of residuals on the same scale.

It may seem odd to include the different residuals on the same plot

- but the data and constraint cases convey distinct information,
- though the two kinds of residuals are linearly related.

AVP for state average expenses per hospital admission



Slope 0.007 (SE 0.002), strongly influenced by Guam (two points at left).

Transforming the outcome or predictors

Transforming y or a predictor can remove curvature or interactions and can make residuals look normally distributed.

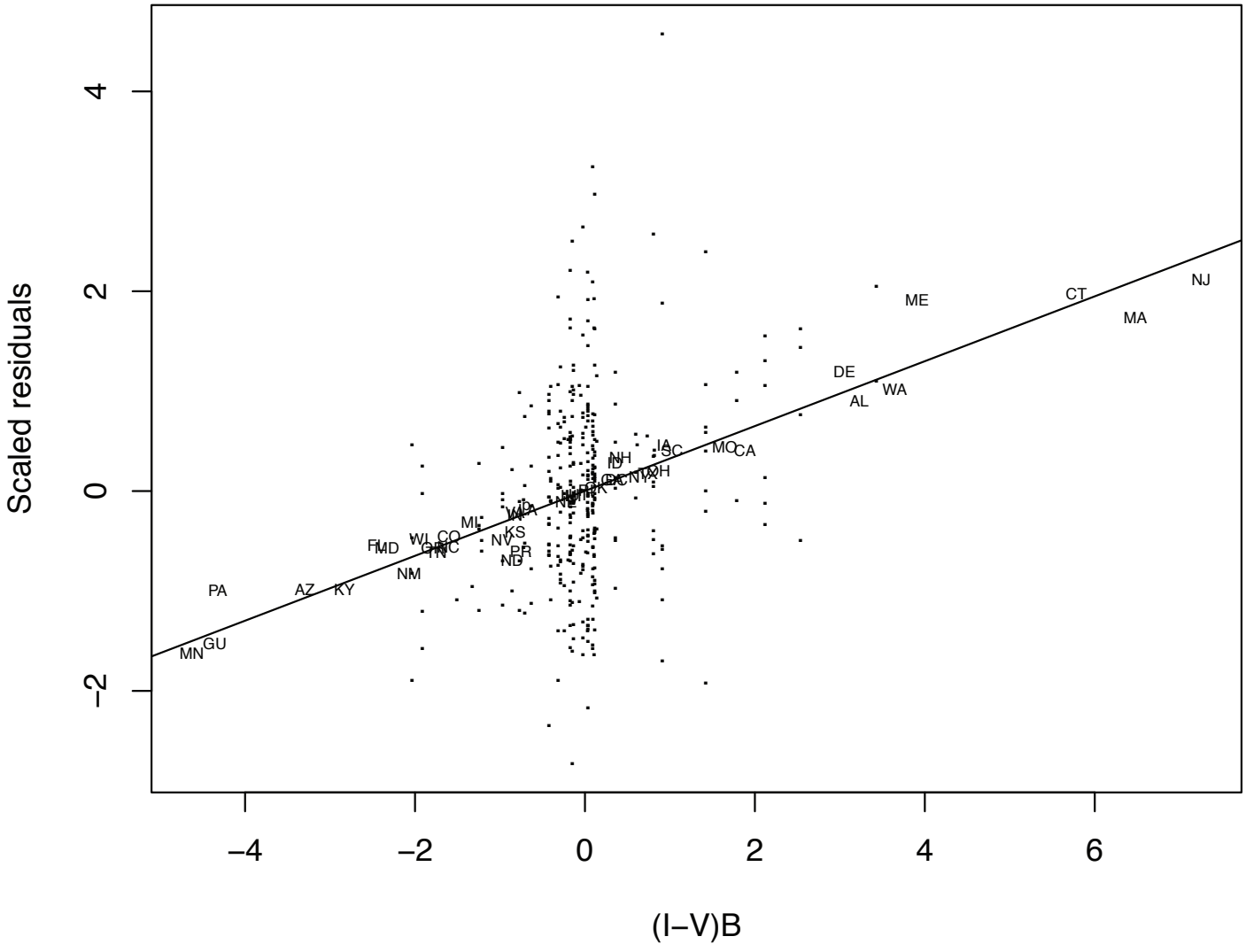
The Box-Cox transformation family for y:

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0. \end{cases}$$

Here's a graphical method to examine the evidence about λ:

- Expand $y^{(\lambda)}$ as a linear Taylor series around $\lambda = 1$.
- $\Rightarrow$ a new predictor $\mathbf{B}$ with coefficient that's linear in λ.
  - Andrews: $\mathbf{B}_I = \hat{y}_I \log(\hat{y}_I) - \hat{y}_I + 1$; $\hat{y}_I$ is the fit for untransformed y.
- Draw the added variable plot for $\mathbf{B}$.

Andrews AVP for the HMO data



OLS fit: slope 0.33 (SE 0.04) $\Rightarrow \lambda = 0.67$ (SE 0.04).

Case influence: Delete a case; how do estimates change?

For the linear model $\mathbf{y} = \mathbf{A}\beta + \epsilon$, there's a handy updating formula.

If (I) indicates deleting cases indexed by I, then

$$(\mathbf{A}'_{(I)}\mathbf{A}_{(I)})^{-1} = (\mathbf{A}'\mathbf{A})^{-1} + (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'_I(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{A}_I(\mathbf{A}'\mathbf{A})^{-1}$$

where $\mathbf{A}_I$ is the deleted rows of $\mathbf{A}$ and $\mathbf{H}_I = \mathbf{A}_I(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'_I$.

If only case I is deleted, and subscript (I) indicates deleting case I:

$$\hat{\beta}_{(I)} - \hat{\beta} = -(\mathbf{A}'\mathbf{A})^{-1}a_I r_I/(1 - h_I)$$

where a_I is the I^{th} row of $\mathbf{A}$, r_I is the I^{th} residual in the full-data fit, $h_I \in (0, 1)$ is the I^{th} diagonal element of $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$.

Case influence measures for simple linear models

Cook's distance measures how $\hat{\beta}$ changes from deleting case I,

$$\begin{aligned} d_I &= (\hat{\beta}_{(I)} - \hat{\beta})' \mathbf{A}' \mathbf{A} (\hat{\beta}_{(I)} - \hat{\beta}) / p\hat{\sigma}^2 \\ &= r_I^2 \frac{h_I}{(1-h_I)^2} \frac{1}{p\hat{\sigma}^2} \end{aligned}$$

where $\mathbf{A}$ has p columns and $\hat{\sigma}^2$ is the full-data error-variance estimate.

$\rightarrow$ d_I is large if both r_I^2 and the leverage h_I are large.

The change in $\hat{\sigma}^2$ from deleting case I:

$$\hat{\sigma}^2_{(I)} - \hat{\sigma}^2 = \frac{1}{n-p-1}(\hat{\sigma}^2 - \frac{r_I^2}{1-h_I}).$$

Under the assumed model, $E(r_I^2) = (1-h_I)\sigma^2$ and $E(r_I) = 0$

$\Rightarrow$ deleting case I reduces $\hat{\sigma}^2$ if r_I^2 is large relative to its variance.

These measures apply to MLM data *and* constraint cases

Objection: Constraint and prior cases are different from data cases.

But any modeling choice inserts information into the analysis.

And it's handy to use one diagnostic for all kinds of cases.

Deleting a prior case = setting the prior variance to infinity.

Deleting a constraint case:

Simple RE model for the HMO dataset: Deleting MN's constraint case implies MN's posterior mean doesn't shrink toward μ. (Boring.)

It may also affect the posteriors of θ_i from other states. (Boring.)

When the state-level model is more complicated, the influence of MN's constraint case on MN's θ_i reflects on the state-level model.

Single-parameter influence measures are better for MLMs

Cook's distance describes case influence on the mean-parameter *vector*.

For an MLM, the mean parameter vector has many elements; a case's influence is often felt strongly by one parameter and weakly by others.

$\Rightarrow$ Cook's distance can miss important effects.

$\Rightarrow$ Use a single-parameter analog to Cook's distance, "relative change":

$$RC(i, l) = \text{Change in parameter } i\text{'s estimate from deleting case } l$$
$$\div \text{ full-data posterior SD of parameter } i$$

Case deletion has non-linear effects, thru the variances

Deleting a case changes the information about $\boldsymbol{\Gamma}$ and thus affects the posterior mean of Θ non-linearly.

If Cook's distance is used with fixed $\boldsymbol{\Gamma}$, it's linear in the omitted case.

This suggests re-running the MCMC for each deleted case, but doing so would be onerous computationally.

$\Rightarrow$ I considered two alternatives:

- Fix $\boldsymbol{\Gamma}$ anyway (i.e., ignore the non-linearity).
- Re-use the MCMC draws using importance-sampling weighting.

Importance sampling the MCMC draws

Label the MCMC sequence by $k = 1, \ldots, m$.

Let $\mathbf{Y}_{(l)}$ be $\mathbf{Y}$ with case l deleted, $\boldsymbol{\Psi} = (\Theta, \boldsymbol{\Gamma})$, and g be a function.

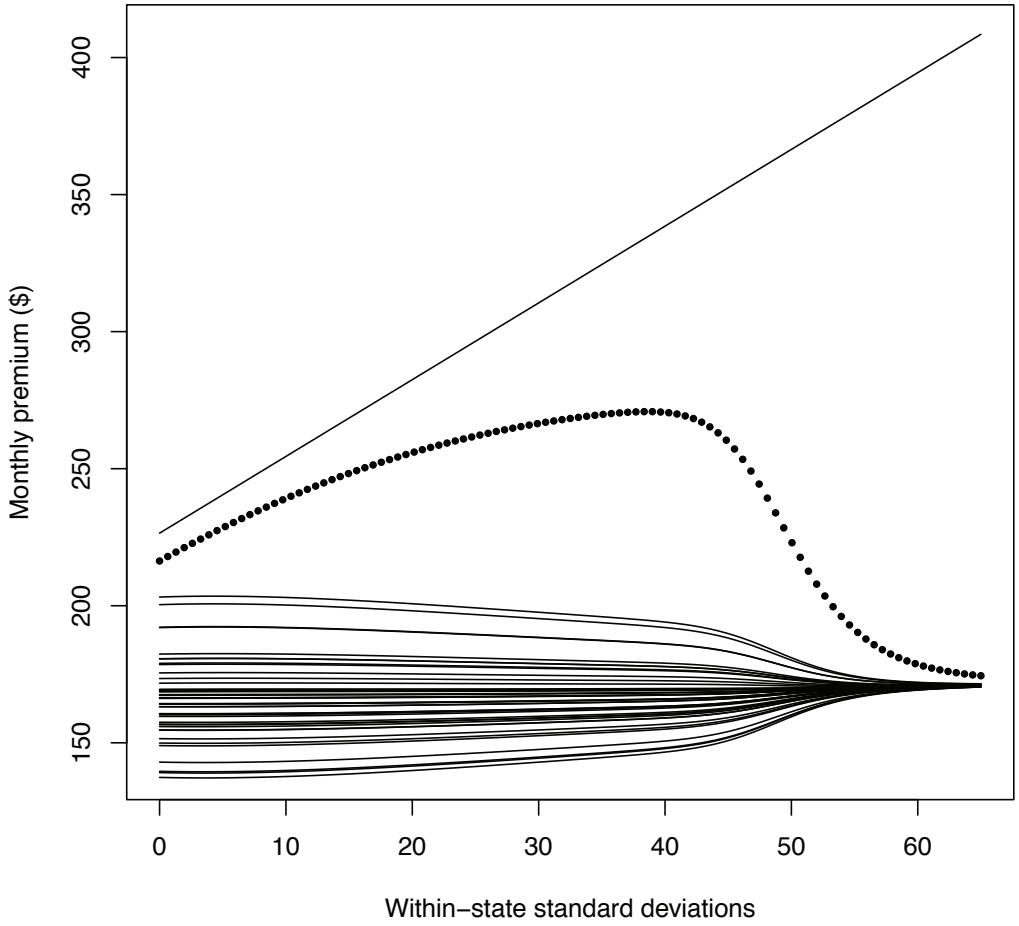
$$\begin{aligned} \text{Then } E(g(\boldsymbol{\Psi})|\mathbf{Y}_{(l)}) &= \int g(\boldsymbol{\Psi}) f(\boldsymbol{\Psi}|\mathbf{Y}_{(l)}) d\boldsymbol{\Psi} \\ &= \int g(\boldsymbol{\Psi}) f(\boldsymbol{\Psi}|\mathbf{Y}) \frac{f(\boldsymbol{\Psi}|\mathbf{Y}_{(l)})}{f(\boldsymbol{\Psi}|\mathbf{Y})} d\boldsymbol{\Psi} \\ &\approx \frac{1}{m} \sum_{k=1}^{m} g(\boldsymbol{\Psi}_k) \frac{f(\boldsymbol{\Psi}_k|\mathbf{Y}_{(l)})}{f(\boldsymbol{\Psi}_k|\mathbf{Y})}, \end{aligned}$$

where $\boldsymbol{\Psi}_k, k = 1, \ldots, m$, is the MCMC sequence.

$f(\boldsymbol{\Psi}_k|\mathbf{Y}_{(l)})/f(\boldsymbol{\Psi}_k|\mathbf{Y})$ is simple, involving one row of $\mathbf{Y} = \mathbf{H}\Theta + \mathbf{E}$.

How exactly case deletion is non-linear

For a simplified version of the HMO dataset ($n_i = 8$), make one of ME's observations an outlier by the amount on the horizontal axis:



Solid line: ME's average; dotted line: ME's posterior mean; bunch of lines: 44 other posterior means.

Fixing Γ (linear approx'n) vs. importance sampling

Both work for cases with small influence, but we care about detecting cases with large influence.

The importance sampler understates case-deletion effects when it fails.

- It reweights the MCMC draws it has; if few are near the case-deleted estimate, importance sampling gives a too-small change.
- It's easy to make this happen in real data.

The linear approximation appears much less sensitive to the non-linearity.

Referee: "It is difficult to imagine an outlier big enough to produce this effect that you couldn't detect by casual examination of the data."

- The previous slide: The weird stuff happens when the outlier is $\approx 45\sigma_e$ from the average of ME's other plans.
- The approximation's fragility depends some on the sample sizes.

The linear approximation applied to the HMO data

For research purposes, I computed the "exact" relative changes (RCs).

For the simple one-way RE model, four constraint cases had $|RC| > 2$ on their own θ_i:

ME (RC 3.4), CT (RC 2.1), PR (RC -2.1), ND (RC -2.4)

The linear approximation was effectively exact.

The importance sampler was not so good:

- For $|RC| < 1$, it was accurate.
- The four biggest $|RC|$ were substantially underestimated; the importance weights were dominated by one MCMC draw.

Residuals, the diagnostic everybody seems to like

For the ordinary linear model $\mathbf{y} = \mathbf{A}\beta + \epsilon$,

- the residuals are $\hat{\epsilon} = (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{y}$;
- $\mathrm{E}(\hat{\epsilon}) = 0$ and $\hat{\epsilon}'\hat{y} = 0$ by construction.

Residuals are plotted

- vs. fitted values to show outliers or non-constant variance;
- vs. candidate predictors: in AVPs; to check for non-linearity; or to check for variance that's a function of the predictor; and
- in quantile plots to check the assumed normal error distribution.

All these uses ignore correlations among the $\hat{\epsilon}$, which are usually small.

Some of these tools don't work in the constraint-case form

For the general model in constraint case form,

$$\left[\begin{array}{c} \mathbf{y} \\ \hline \mathbf{0} \\ \hline \mathbf{M} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{K} & \mathbf{0} \\ \hline \mathbf{S}_1 & \mathbf{S}_2 \\ \hline \mathbf{W}_1 & \mathbf{W}_2 \end{array}\right] \left[\begin{array}{c} \Theta_1 \\ \hline \Theta_2 \end{array}\right] + \left[\begin{array}{c} \epsilon \\ \hline \delta \\ \hline \xi \end{array}\right],$$

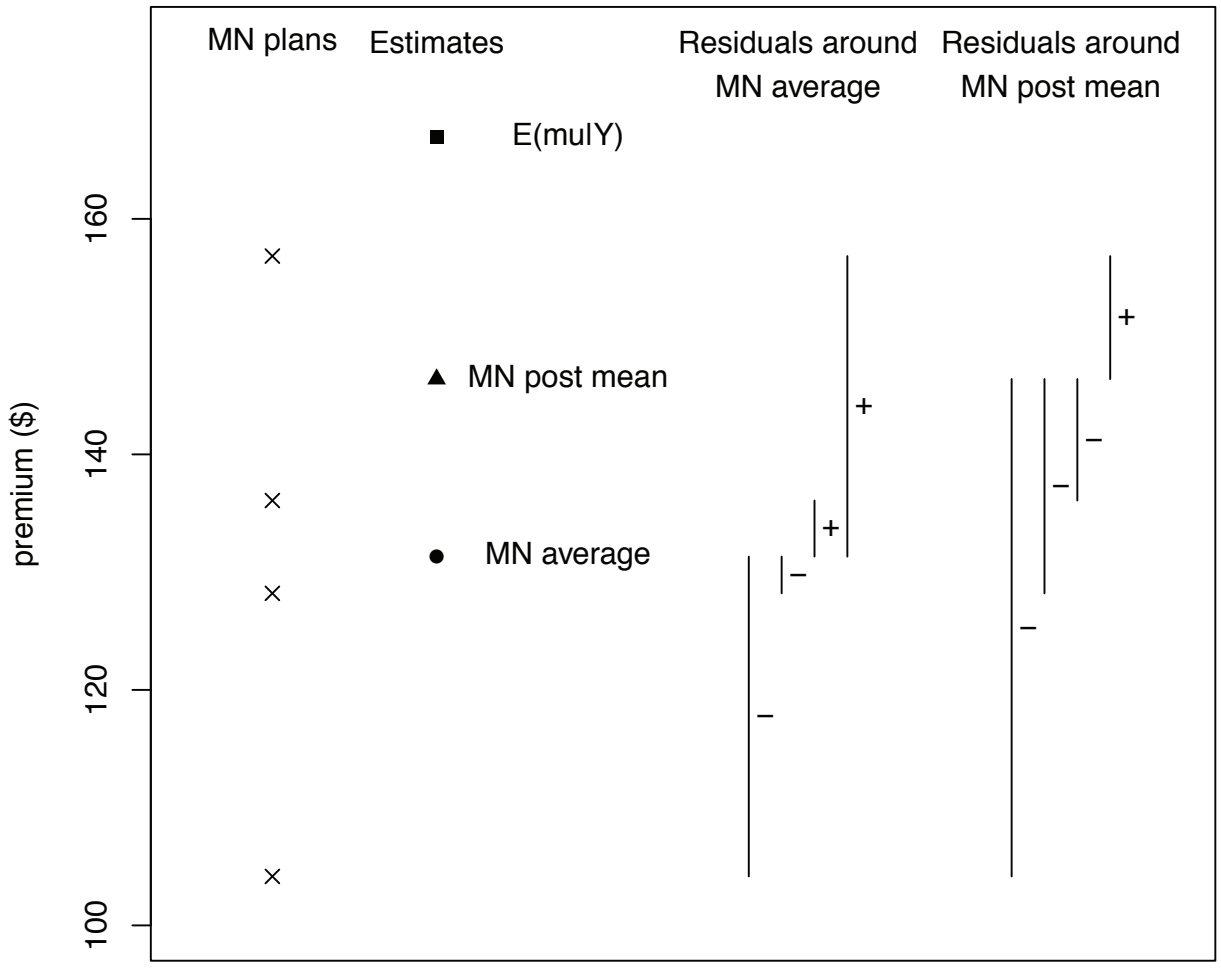
the biggest problem is that for the constraint cases:

residuals: $0 - (\mathbf{S}_1\hat{\Theta}_1 + \mathbf{S}_2\hat{\Theta}_2)$
fitted values: $\mathbf{S}_1\hat{\Theta}_1 + \mathbf{S}_2\hat{\Theta}_2$.

These *are* the correct residuals for this plot;
the correct fitted values are $\mathbf{S}_2\hat{\Theta}_2$.

$\Rightarrow$ Unlike the previous three diagnostics, data and constraint cases need separate residual plots.

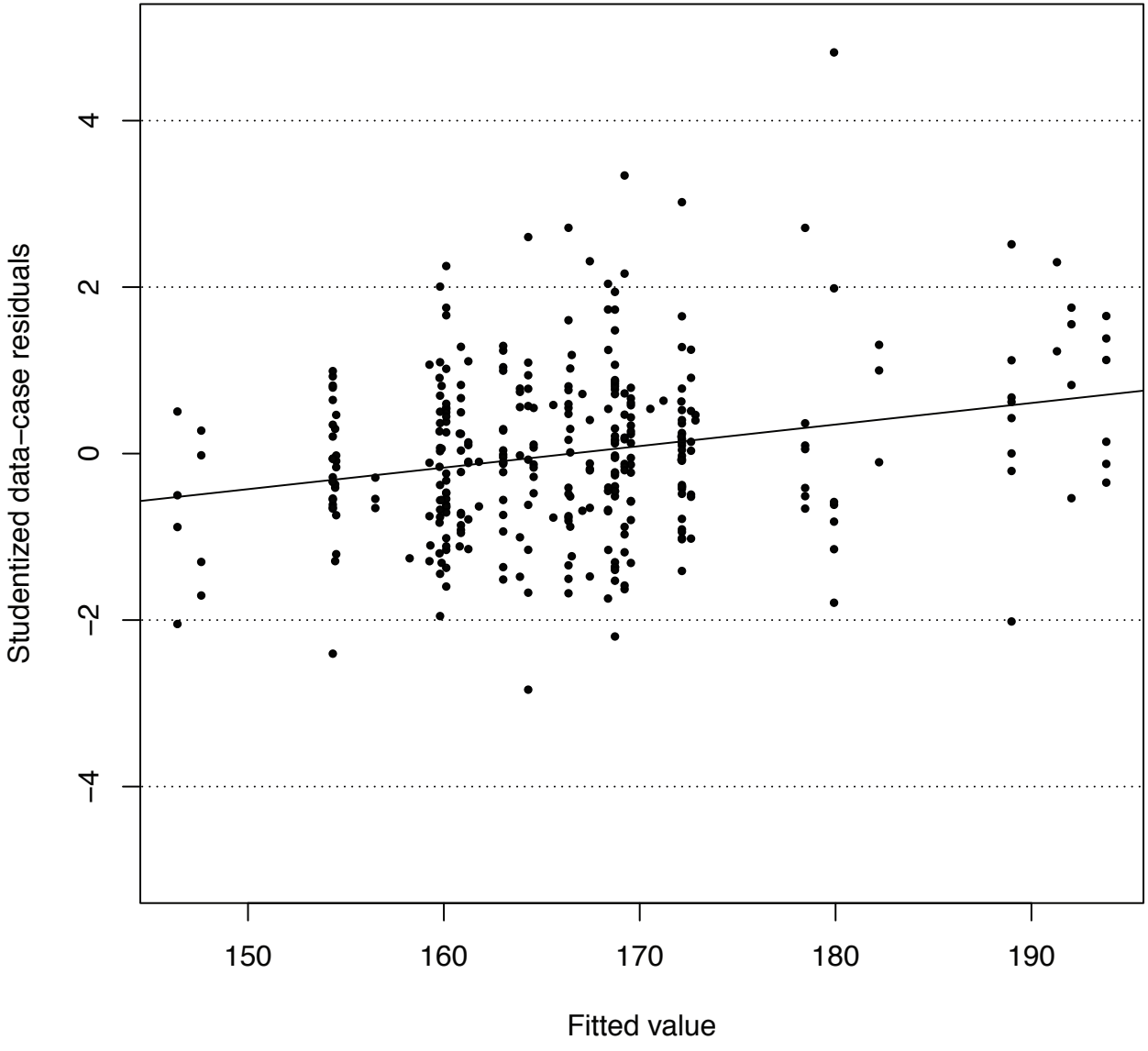
A different unhappy thing happens for data-case residuals



For this model & others, residuals & fitted values are positively correlated.

Here's how this looks in the HMO dataset

The solid line is the LS fit of studentized residuals on fitted values.



Residuals are biased and $\hat{\epsilon}'\hat{y} \neq 0$; what to do?

In MLMs generally, the fit is biased $\Rightarrow$ residuals are biased.

- There's a tidy formula for the bias (RWC; Hodges 2013, Sec. 8.5).
- E.g., penalized splines: the fit is low at peaks, high in valleys.

Hilden-Minton (1995):

- Fit an unshrunk model and use the residuals from that.
- Construct linear transforms of the data-case residuals, $l'\hat{E}_d$ that are independent of each other and unbiased.

LNP (2006):

- Plot the residuals from the full fit, $\hat{e}_l = y_l - X_l\hat{\boldsymbol{\beta}} - Z_l\hat{\mathbf{u}}$, vs. the fitted *fixed effects* $X_l\hat{\boldsymbol{\beta}}$.
- Their result isn't true in the generality they claim.
- For many RPMs, $X_i\hat{\boldsymbol{\beta}}$ is almost completely divorced from the fitted values of interest, e.g., penalized spline.

Hodges view: There's not much you *can* do

Probably the best solution is to try to ignore this structure in residuals.

- For ordinary linear models, we ignore correlations among the residuals.

It's *possible* to estimate the bias and remove it.

- Zhao, Hodges, Carlin (2017), in network meta-analysis, estimating inconsistency of direct and indirect evidence.

It's not clear that this is a good idea in general

- Cunanan, Dai (2014, 2016 projects): Estimating bias of a penalized spline fit works poorly for constructing point-wise CIs.