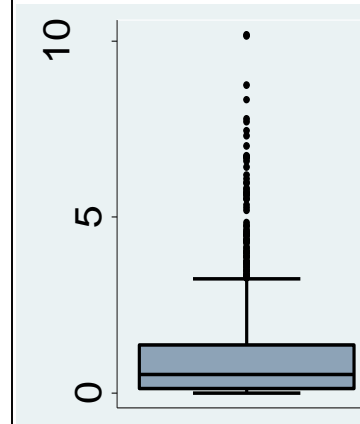


Flagging extreme clusters in mixed models using weighted and self-calibrated predictors

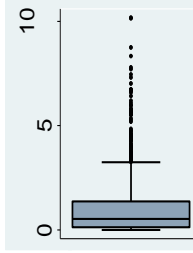
***Charles E. McCulloch,
Division of Biostatistics,
Department of Epidemiology and
Biostatistics
University of California, San Francisco***



December, 2025

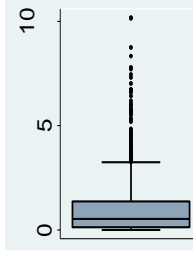
Collaborators

- Joint work with Drs. John Neuhaus and Ross Boylan



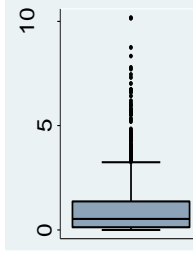
Outline

- Examples of flagging
- Briefly: shrinkage prediction
- Improvement with weighted predictors
- Flagging extreme values
- Poor performance of currently used rules
- Self-calibration
- Numerical comparisons
- Back to asthma example
- Summary

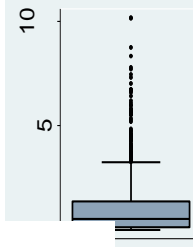


Outline

- Examples of flagging
- Briefly: shrinkage prediction
- Improvement with weighted predictors
- Flagging extreme values
- Poor performance of currently used rules
- Self-calibration
- Numerical comparisons
- Back to asthma example
- Summary

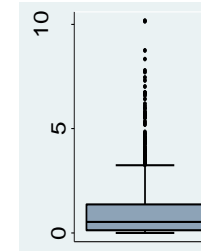


Asthma example



- Are certain regions in the state of California doing a poor job in treating pediatric asthma patients? Follow patients who have been admitted to the emergency room due to asthma.
- Goal: Predict or flag zip codes (postal codes) with high emergency department readmission rates for children.

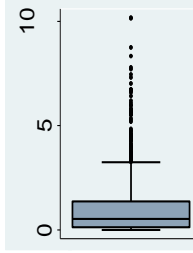
Other examples



- Suppose you are a woman scheduling a C-section. Which hospitals to choose or avoid?
- CA compare: <https://calhospitalcompare.org>
- Risk-adjusted operative mortality rates for coronary artery bypass surgery for each of 276 surgeons in California in 2017-2018, only 3 surgeons are flagged as above average.
- Cross-over trial for treatment of urinary problems in elderly men: which are not benefiting from drug and can be de-prescribed?

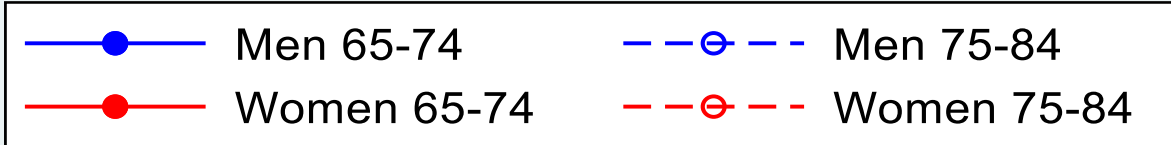
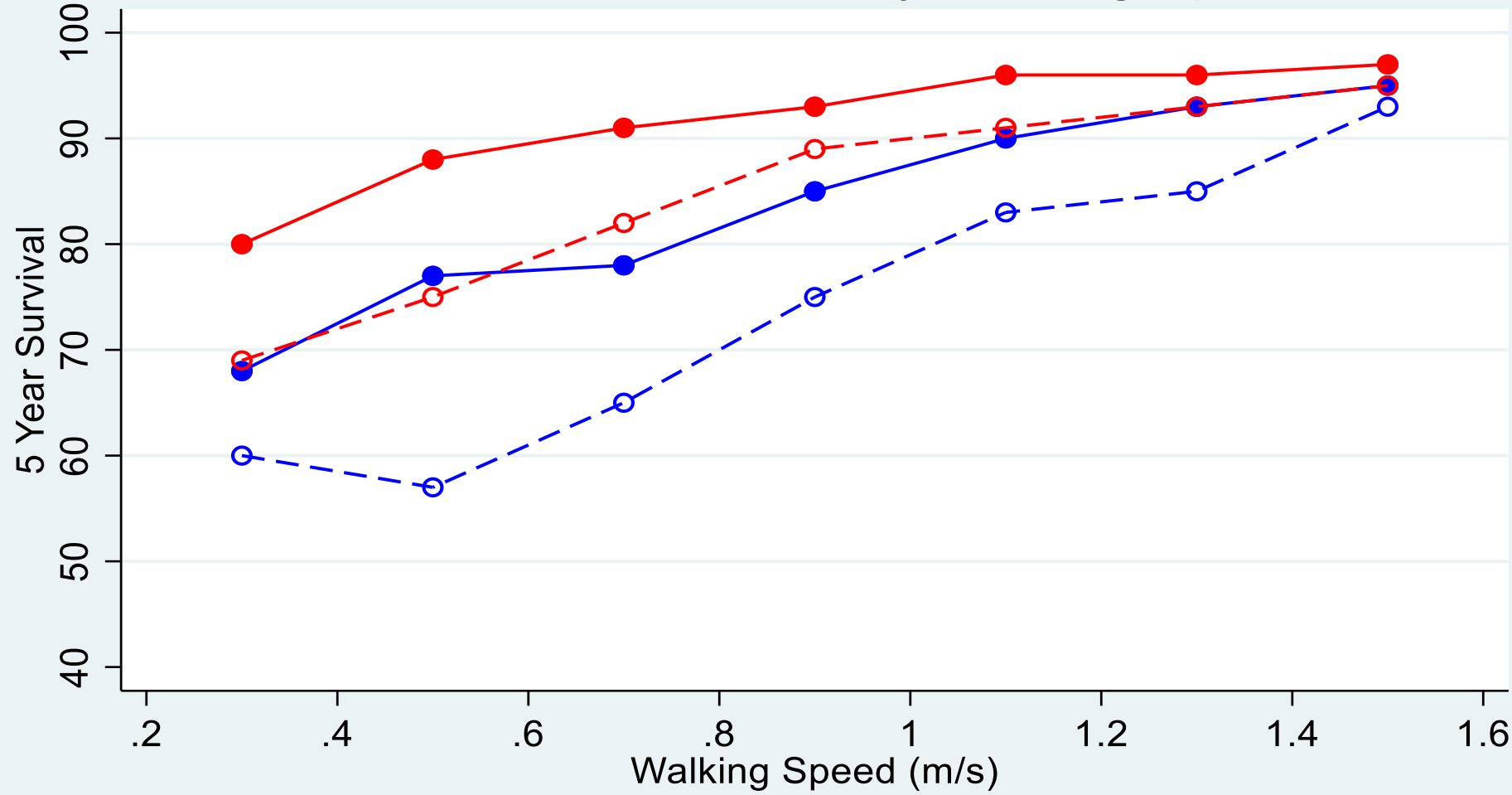
Outline

- Examples of flagging
- **Briefly: shrinkage prediction**
- Improvement with weighted predictors
- Flagging extreme values
- Poor performance of currently used rules
- Self-calibration
- Numerical comparisons
- Back to asthma example
- Summary



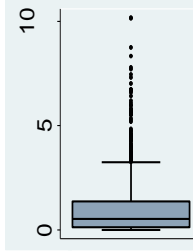
Gait Speed and Survival in Older Adults

5 Year Survival Rates by Walking Speed



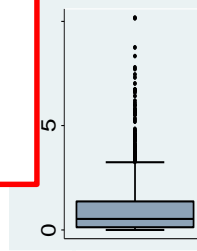
OAI

- Osteoarthritis Initiative
<https://nda.nih.gov/oai/about-oai>
- Designed as a study of natural history of knee arthritis. Approximately 5,000 participants followed longitudinally.
- What variables are associated with slow walking speed, e.g., sex or age of the participant?
- Also, use the first 4 yearly visits to predict the average value of walking speed in 3 later visits.



Some sample data

Considerable person to person variation in walking speed

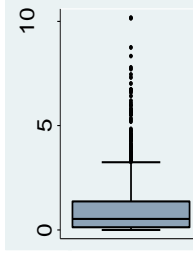


Person	Visit			
	1	2	3	4
1	0.93	0.93	0.81	0.71
2	1.17	1.33	1.25	1.27
3	1.59	1.65	1.65	1.72
4	1.39	1.38	1.14	1.18
5	1.77	1.71	1.82	1.71
6	1.67	1.80	1.60	1.74

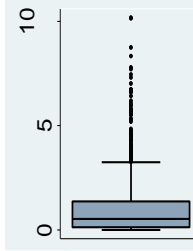
Need to accommodate person to person variation in the analysis. Mixed models allow person-specific terms, e.g., person-specific intercepts.

Outline

- Introduction: OAI study
- **Introduction: mixed models**
- Introduction: shrinkage predictors
- Improvement with weighted predictors
- Prediction accuracy for extreme values
- Flagging extreme values
- Self-calibration
- Back to asthma example
- Summary



A simple mixed model



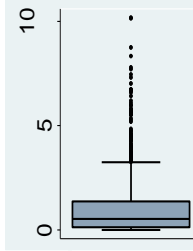
- Each person has a conceptual average value – the average speed if you got to take an infinite number of measurements. Denote the average for person i as μ_i .
- There is variation in speed within a person and (likely) between people.

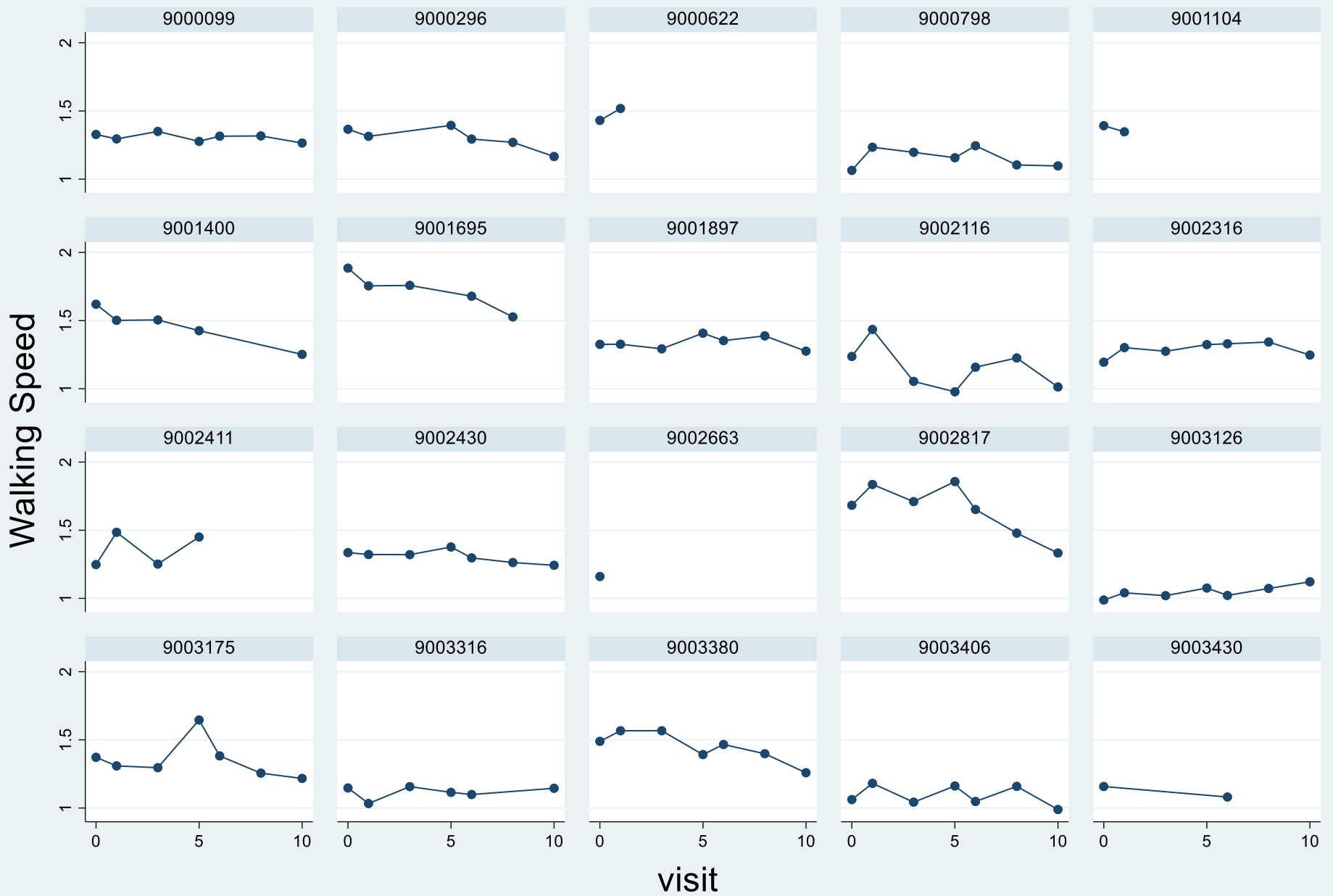
$$Y_{ij} = WS_{ij} = \mu_i + \varepsilon_{ij}$$

$$E[\mu_i] = \mu, \text{var}(\mu_i) = \sigma_\mu^2, \text{var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$$

A simple mixed model

- You might consider treating person as a categorical variable.
- But that will not work if we are interested in things like the difference between men and women (the categorical person predictor would override the sex effect).
- Potentially inefficient to fit a 5,000 level categorical variable. And ill advised with a binary outcome.
- Instead assume the person-specific terms follow a normal distribution (usually innocuous).





Graphs by ID

Some sample data

Person	Visit				Average Visits 1-4	Average Visits 5-7
	1	2	3	4		
1	0.93	0.93	0.81	0.71	0.84	
2	1.17	1.33	1.25	1.27	1.26	
3	1.59	1.65	1.65	1.72	1.56	
4	1.39	1.38	1.14	1.18	1.27	
5	1.77	1.71	1.82	1.71	1.75	
6	1.67	1.80	1.60	1.74	1.70	

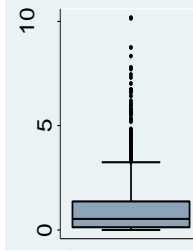
Use this to predict the average of visits 5-7?

Note: Even though there is a slight decline in walking speed with visits, it is negligible (about 0.01 over the course of the study, on average).

Can we do better?

- Instead of using the sample mean for each person from visits 1 to 4 to predict, consider a minor variation, a linear combination of the sample mean for the i^{th} cluster (person):

$$a + b\bar{Y}_i.$$

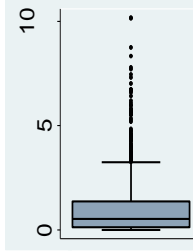


Can we do better?

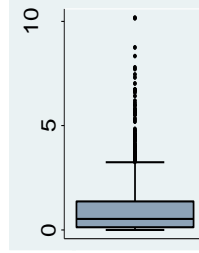
- Want to derive a predictor of μ_i . What predictor has the lowest mean square error of prediction (MSEP) across all the people?

$$\min_{a, b} E[(a + b\bar{Y}_i. - \mu_i)^2]$$

- Quadratic in a and b so just take derivatives and set equal to 0.



Shrinkage predictor



- Plugging back in, predictor for μ_i is:

Overall mean

$$\mu + \frac{\sigma_{\mu}^2}{\sigma_{\mu}^2 + \frac{\sigma_{\epsilon}^2}{n_i}} (\bar{Y}_{i\cdot} - \mu)$$

Deviation from overall mean

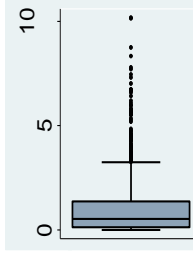
- Predict overall mean plus a proportion (<1) of the deviation from the overall mean. “Shrinks” to the overall mean.
- Made virtually no assumptions.

Shrinkage predictor

- Behavior of the shrinkage term:

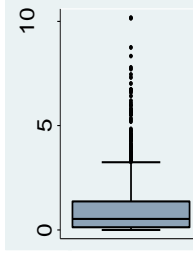
$$b = \frac{\sigma_{\mu}^2}{\sigma_{\mu}^2 + \frac{\sigma_{\epsilon}^2}{n_i}}$$

- Between 0 and 1.
- Close to 1 if n_i is large or noise is small (small variance for ϵ) \rightarrow predict cluster specific mean when the estimate is precise.
- Close to 0 if no variation between clusters \rightarrow predict overall mean when all clusters seem the same.
- *BLUP* : “best linear unbiased predictor”



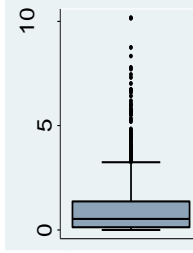
Software for mixed models

- Stata `meq` `glm` R `glmer` SAS `glimmix`
- Or specialized routines for different outcome types
 - Numeric (`mixed` in Stata, `lmer` in R, `mixed` in SAS)
 - Binary (`melogit`, `meprobit` in Stata)
 - Count (`mepoisson` in Stata)
 - Ordered categorical (`meologit`, `meoprobit` in Stata)
- *All of these can produce shrinkage predictors, e.g., the BLUP for linear mixed models.*



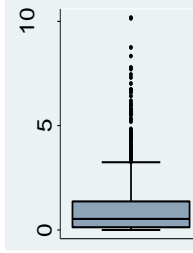
Outline

- Examples of flagging
- Briefly: shrinkage prediction
- **Improvement with weighted predictors**
- Flagging extreme values
- Poor performance of currently used rules
- Self-calibration
- Numerical comparisons
- Back to asthma example
- Summary

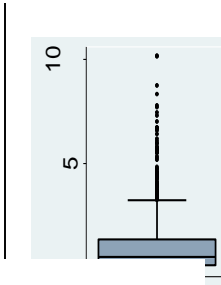


How to get at extremes?

- Has been suggested to assume heavy-tailed distributions, such as a t distribution with low d.f.
- Because a t distribution with low d.f. is heavy-tailed compared to a normal distribution, it might accommodate extreme values without much shrinkage.
- Kalbfleisch and Wolfe (2013) suggested fixed effects (no shrinkage).



Model



Generally conducted in the context of a mixed model.

Base model for the special case for a numeric outcome, random intercepts, u_i , and normality assumptions and covariates x :

$$Y_{ij} = u_i + x'_{ij}\beta + \varepsilon_{ij}$$

$$u_i \sim N(0, \sigma_u^2) \perp \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$Y_{ij} = \sigma_u z_i + x'_{ij}\beta + \varepsilon_{ij}$$

How to emphasize extremes?

Specify a weight function $w(z)$ that more heavily weights extremes. Optimal predictor:

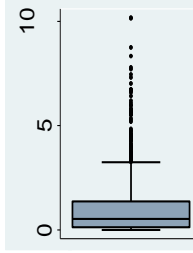
$$\min_{\tilde{z}} E[(\tilde{z} - z)^2 w(z)]$$

Solution given by

$$\tilde{z}_w = E[zw(z)|Y]/E[w(z)|Y]$$

In words: the average value of the random effect weighted by the weight function, normalizing the weights and conditioning on the observed data.

McCulloch & Neuhaus (2021, JASA)



Relation to the best predictor

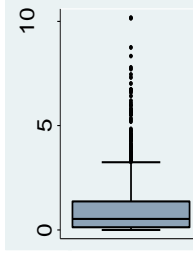
When the weight function is constant, e.g., $w(z)=1$ this reduces to the minimum MSE predictor:

$$\min_{\tilde{z}} E[(\tilde{z} - z)^2]$$

Solution given by the Best Predictor

$$\tilde{z}_{BP} = E[z|Y]$$

For our linear mixed model with normality assumptions, that is the shrinkage predictor noted earlier.



Notes and notation

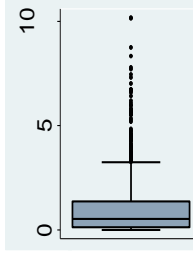
- Recall we are dealing with a standardized version of the random effect, $u_i/\sigma_u = z_i$.
- \tilde{Z}_{BP} “usual” best predictor. \tilde{Z}_{FX} “fixed effects” predictor (observed, standardized deviation).
 - $\tilde{Z}_{FX} = (\bar{Y}_i - \bar{x}'_i \beta) / \sigma_u$
 - $\tilde{Z}_{BP} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2 / n_i} \tilde{Z}_{FX} = \frac{1}{1 + R_i} \tilde{Z}_{FX}$
- Results depend on variances & n only through $R_i = \frac{\sigma_\varepsilon^2 / n_i}{\sigma_u^2}$.
- R_i large ~ noisy data and/or small separation.

$$w(\mathbf{z}) = \exp(\lambda \mathbf{z}^2)$$

One example of a weighted predictor is exponential Squared weighting. λ controls emphasis on extremes. Optimal predictor is given by

$$\frac{1}{1 + R_i(1 - 2\lambda)} \frac{\bar{Y}_{i\cdot} - \bar{x}'_{i\cdot} \beta}{\sigma_u} = \frac{1}{1 + R_i(1 - 2\lambda)} \tilde{Z}_{FX}.$$

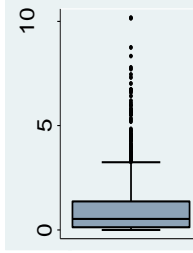
As λ varies from 0 to 0.5 (its maximum), \tilde{Z}_{SQ} varies between \tilde{Z}_{BP} (usual shrinkage) up to \tilde{Z}_{FX} (no shrinkage). So, it gives intermediate values as λ varies.



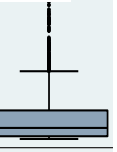
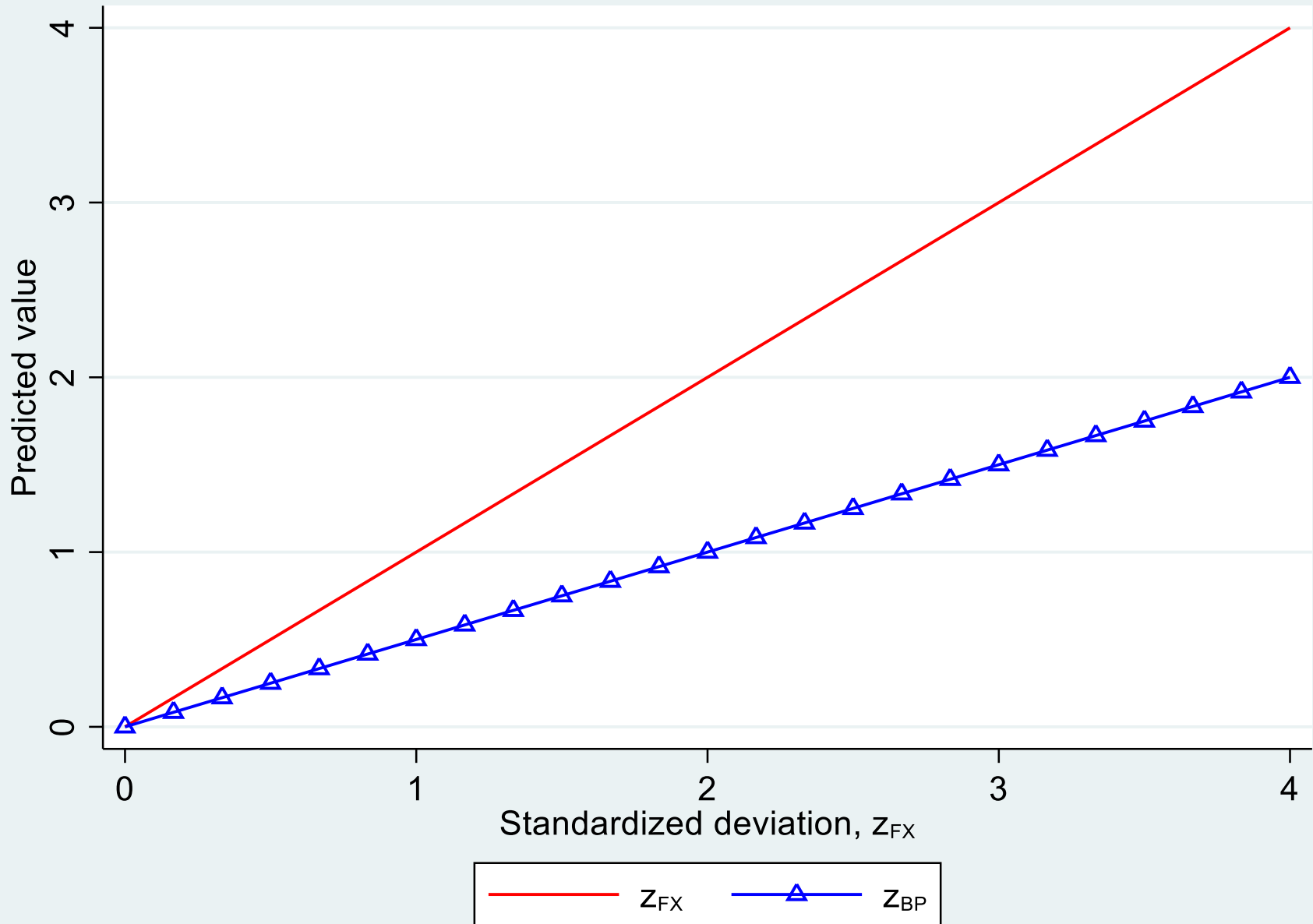
$$w(\mathbf{z}) = \exp(\lambda \|\mathbf{z}\|)$$

Another example is exponential ABsolute weighting.

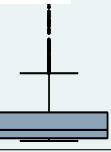
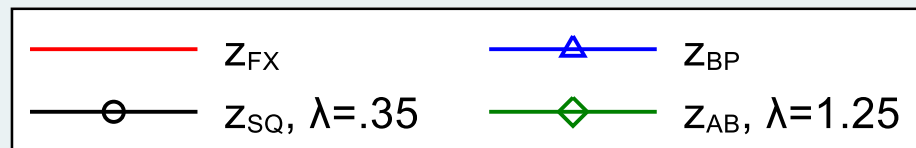
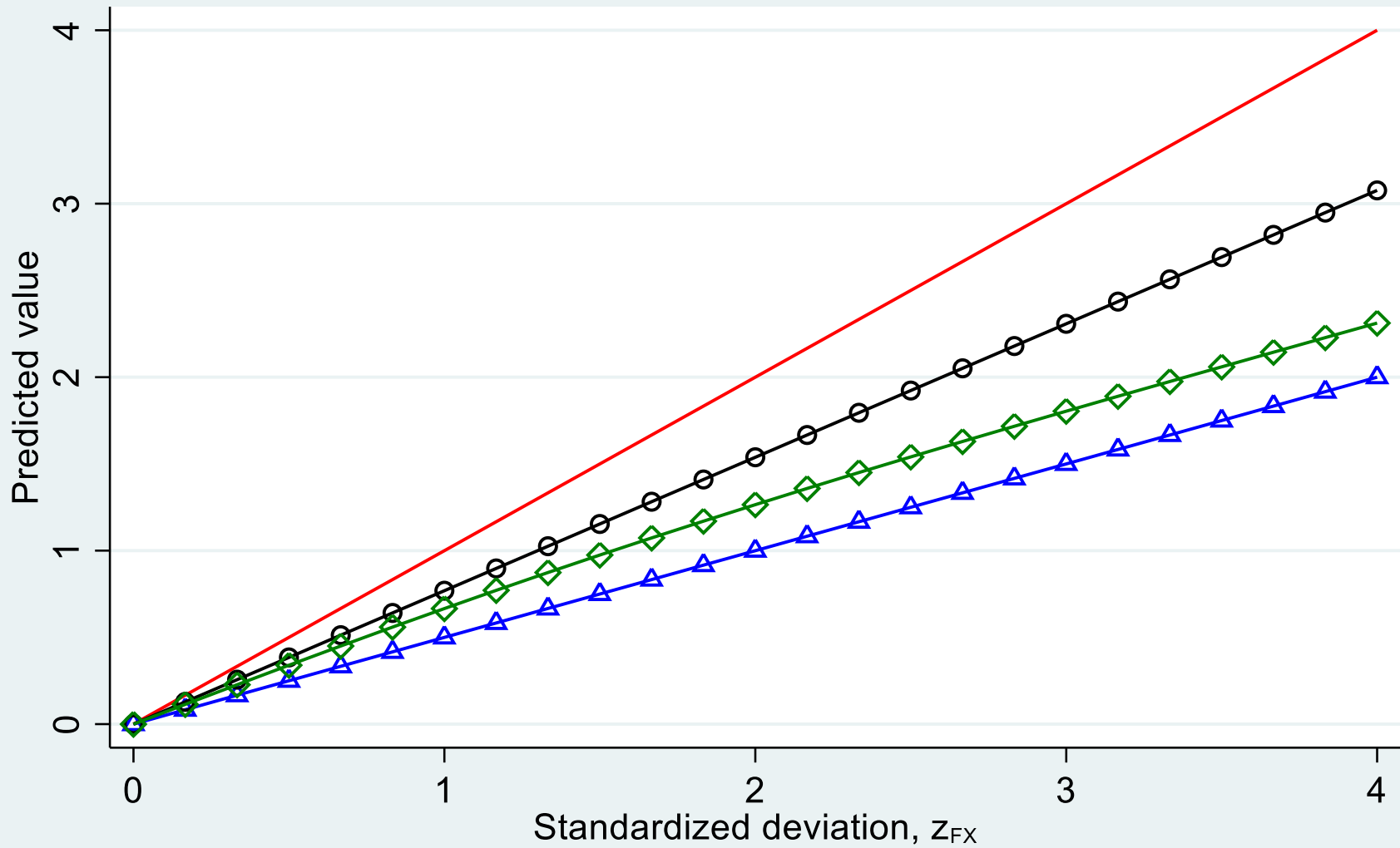
We denote that optimal predictor by \tilde{z}_{AB} .



Comparison of shrinkage, R=1

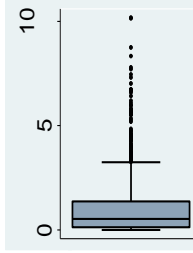


Comparison of shrinkage, R=1

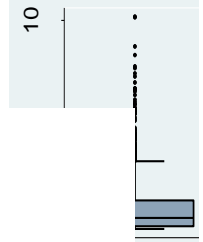


Outline

- Examples of flagging
- Briefly: shrinkage prediction
- Improvement with weighted predictors
- **Flagging extreme values**
- Poor performance of currently used rules
- Self-calibration
- Numerical comparisons
- Back to asthma example
- Summary



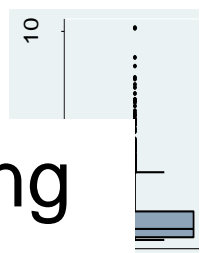
Flagging high values of z



- We want to flag a cluster (declare it is extreme) if it *is* extreme. For example, greater than a threshold, τ , e.g., $z > \tau$ with $\tau = 1.645$.
- Use a predictor, \tilde{z} and its accuracy, $\tilde{\sigma}$.
- Usual form of flagging rule treats it like a one-sided hypothesis test
- Flag a cluster as high if

$$\tilde{z} - z_{\delta} \tilde{\sigma} > \tau \quad \text{or} \quad \frac{\tilde{z} - \tau}{\tilde{\sigma}} > z_{\delta}$$

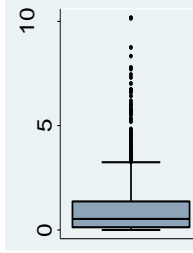
Flagging high values of z

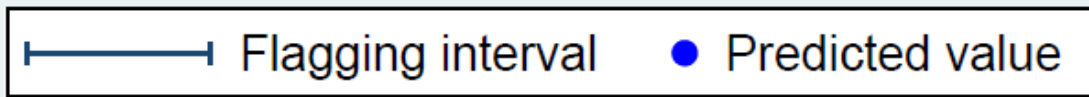
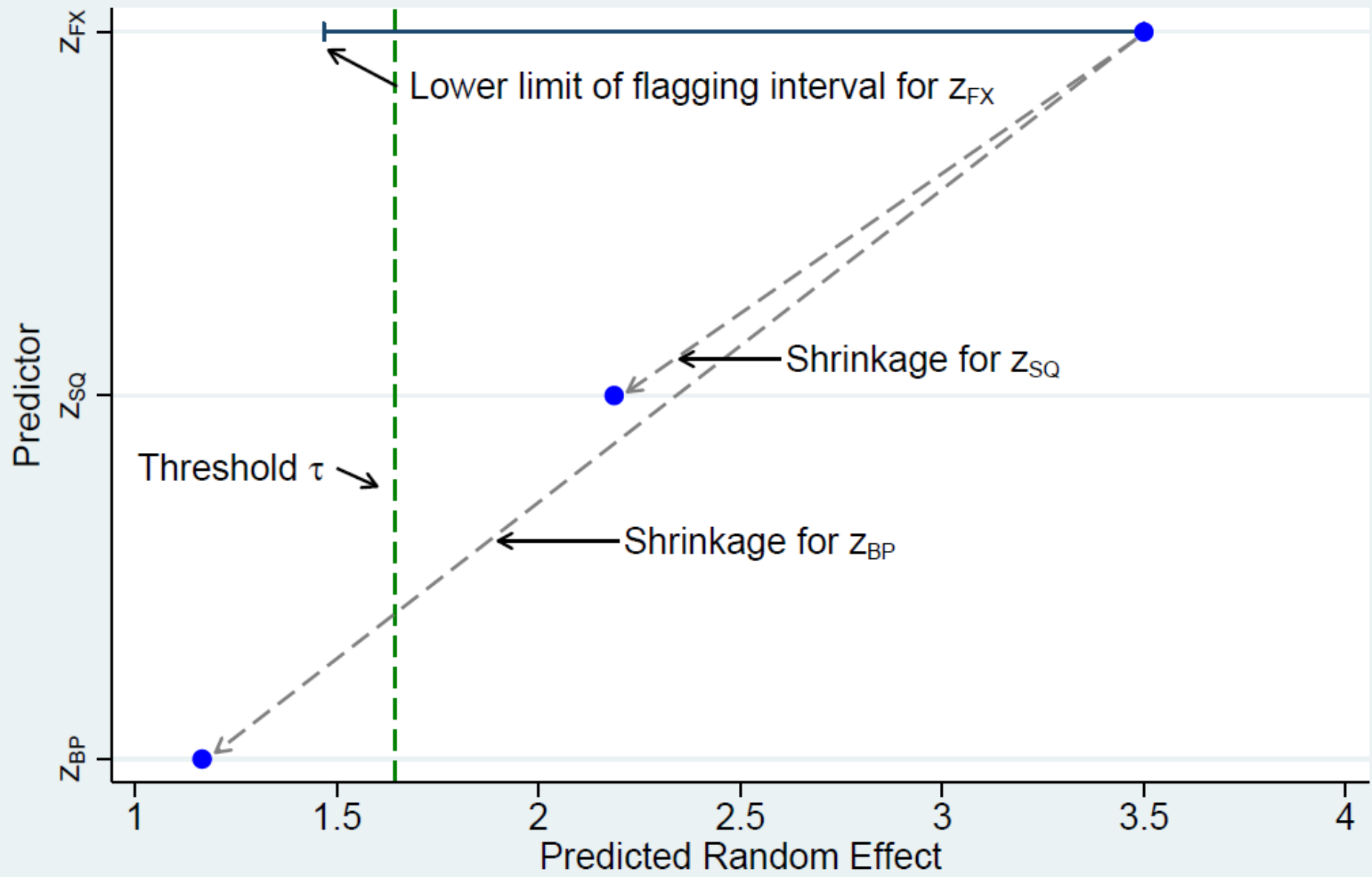


- Because \tilde{z}_{BP} shrinks, it is ill-suited for finding extreme values.
- If you further use a rule like a one-sided hypothesis test (e.g., $\tilde{z}_{BP} - 1.645 * SE > \tau$) the probability of flagging is way too low.
- Reformulate: control probability of incorrect flagging (flag if $z < \tau$) to α or less, maximize probability of correct flagging (flag if $z > \tau$) .

Outline

- Examples of flagging
- Briefly: shrinkage prediction
- Improvement with weighted predictors
- Flagging extreme values
- Poor performance of currently used rules
- **Self-calibration**
- Numerical comparisons
- Back to asthma example
- Summary





Self-calibration: definition

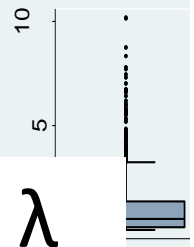
For a given predictor, if we can find a value of λ (depending on α , R and τ) to satisfy:

$$\Pr\{\tilde{z}_\lambda > \tau \mid z < \tau\} = \alpha$$

we call it self-calibrated.

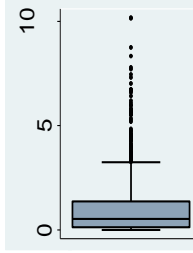
If we can find such a rule it is very convenient to use. First, it controls the incorrect flagging rate to α . Second, we flag a cluster as high whenever its predicted value exceeds τ .

No need for a hypothesis testing formulation.



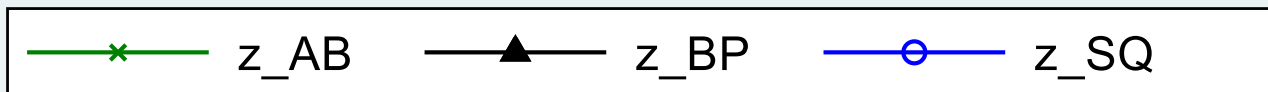
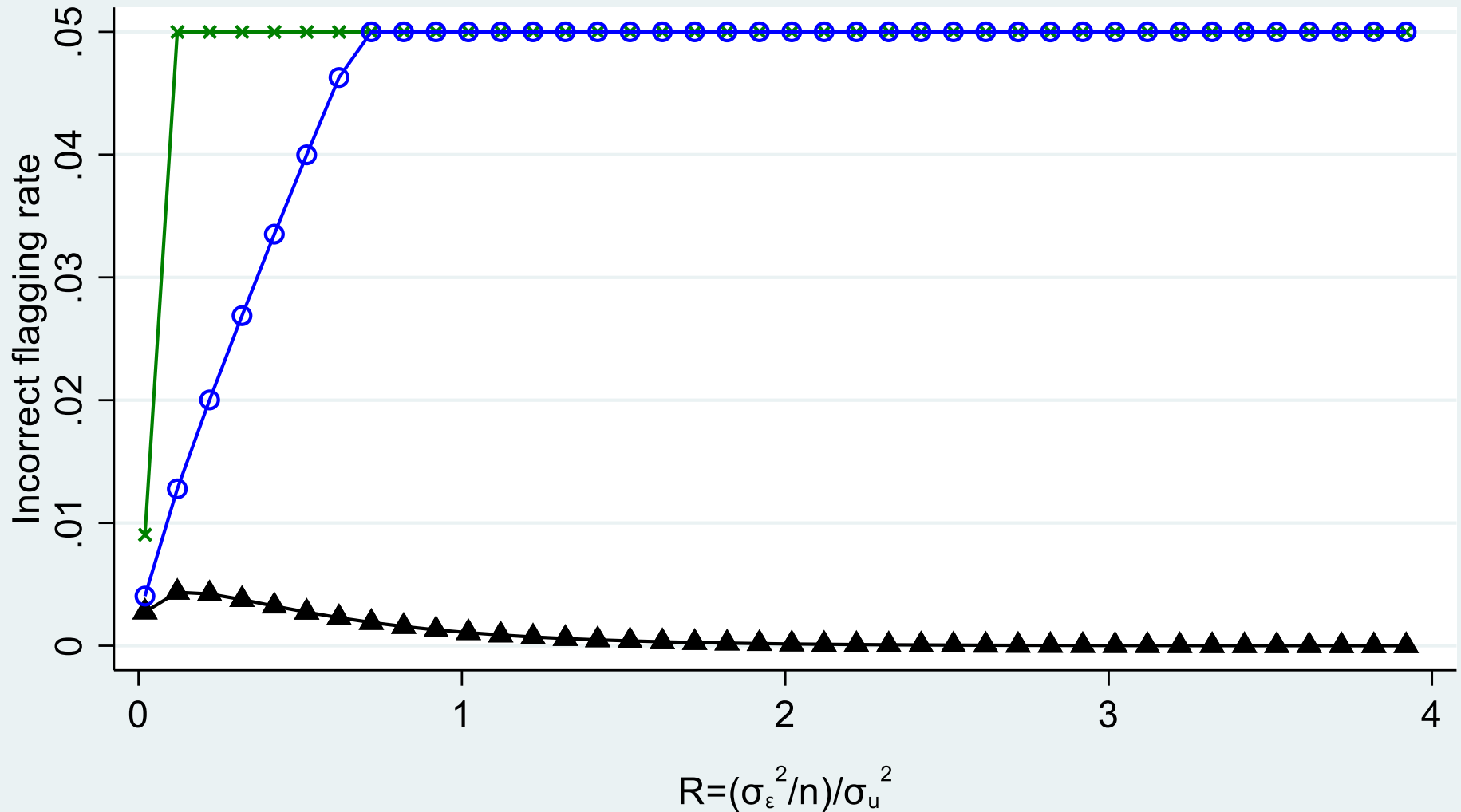
Outline

- Examples of flagging
- Briefly: shrinkage prediction
- Improvement with weighted predictors
- Flagging extreme values
- Poor performance of currently used rules
- Self-calibration
- **Numerical comparisons**
- Back to asthma example
- Summary



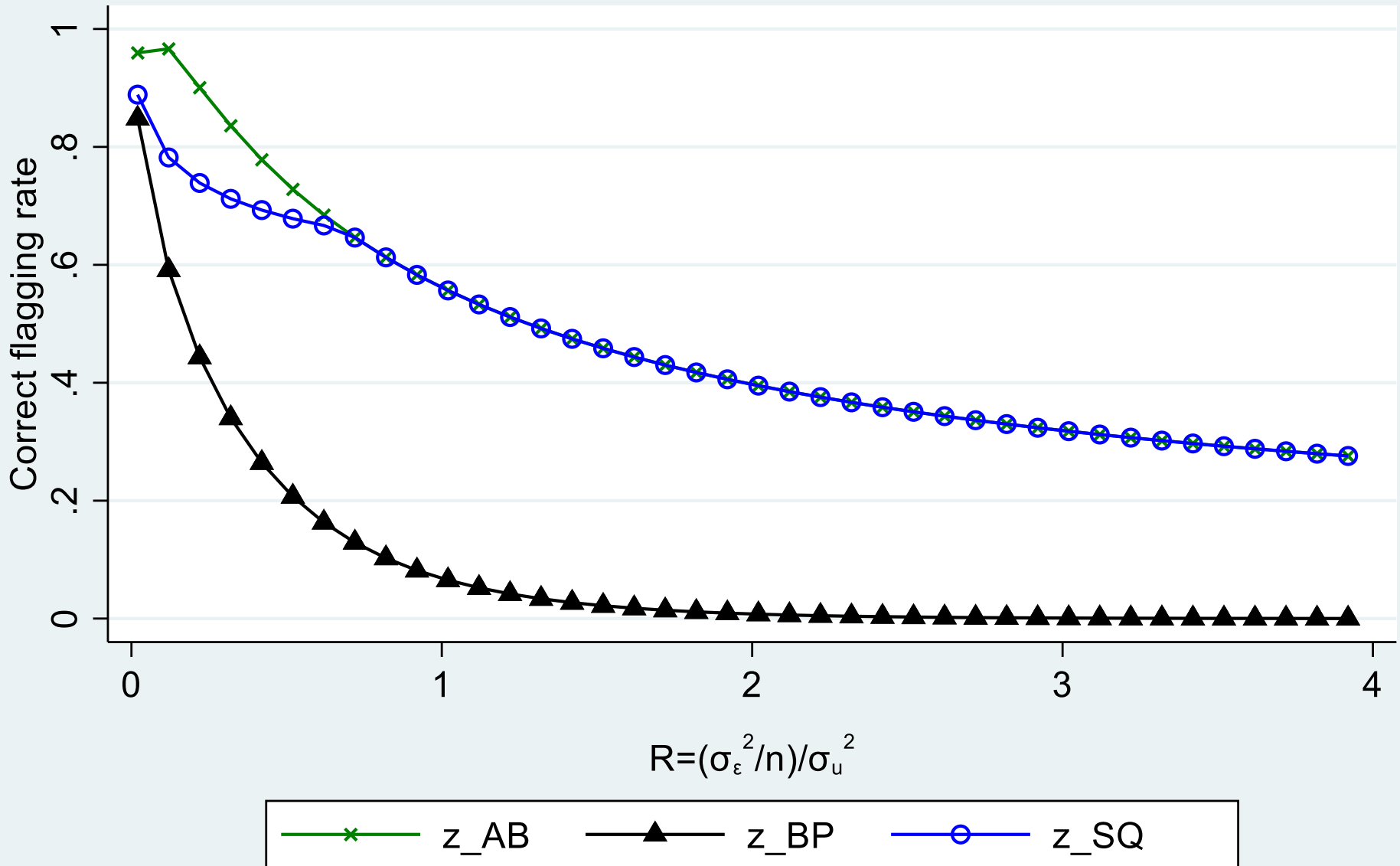
Incorrect flagging rates for self-calibrated predictors

For α of .05 and τ of 1.96



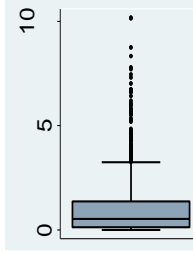
Correct flagging rates for self-calibrated predictors

For α of .05 and τ of 1.96

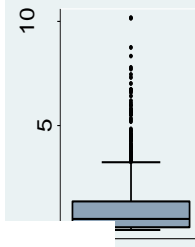


Outline

- Examples of flagging
- Briefly: shrinkage prediction
- Improvement with weighted predictors
- Flagging extreme values
- Poor performance of currently used rules
- Self-calibration
- Numerical comparisons
- **Back to asthma example**
- Summary

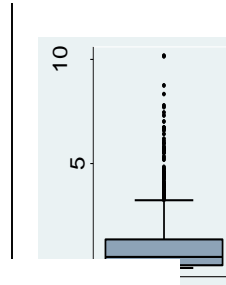


Asthma example



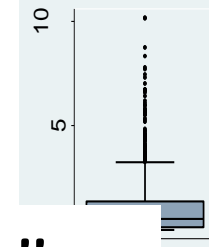
- Goal: Predict or flag zip codes with high emergency department readmission rates.
- Partition data into training and validation.
- Call zip codes in the validation data in top 10% of observed proportions of readmissions “extreme.” Remaining 90% “non-extreme.”
 - Restrict to zip codes with $n_i \geq 100$ for stability of determination of extreme.

Asthma example: analysis



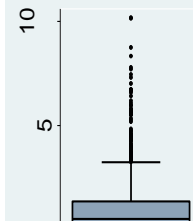
- Objective: flag zip codes with $z > 1.28$, i.e., top 10% of readmission random effects.
- Allow $\alpha = 0.10$ incorrect flagging rate.
- Calculate \tilde{Z}_{BP} and the self-calibrated versions of \tilde{Z}_{SQ} and \tilde{Z}_{AB} .
- Flag zip codes if $\tilde{z} > 1.28 = \tau$
- Compare flagged zip codes to “extreme” zip codes from validation sample.

Asthma example: results



- \tilde{Z}_{BP} incorrectly flags 1 of 223 “non-extreme” zip codes (0.4%), far below the nominal rate of 10%.
- It correctly flags only 1 of 25 “extreme” zip codes.
- Both \tilde{Z}_{SQ} and \tilde{Z}_{AB} incorrectly flag 21 of 223 “non-extreme” zip codes (9.4%), close to the nominal rate of 10%.
- They correctly flag 8 of 25 “extreme” zip codes.

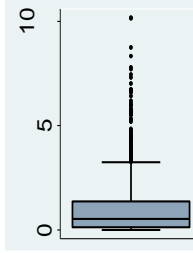
Asthma example: an “oracle” analysis



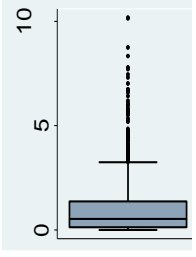
- How well could we do with complete information? The classification of extreme zip codes in the validation sample is fallible.
- Simulate z_i and data with the same μ , σ , n_i as our data.
- We know which z_i are truly in the top 10%. Flag those clusters.
- On average, the oracle rule flags 8.5 of 25 “extreme” zip codes.

Summary

- If interest focuses on extreme clusters, weighted predictors have lower (sometimes much lower) MSEP.
- Traditionally used flagging rules based on \tilde{Z}_{BP} or \tilde{Z}_{FX} (results not shown) perform very poorly.
- Very simple flagging rules based on weighted predictors, \tilde{Z}_{SQ} and \tilde{Z}_{AB} , control the incorrect flagging rate and have much higher correct flagging rates.
- \tilde{Z}_{AB} works somewhat better than \tilde{Z}_{SQ} .



Questions?



Some Stata code

Linear mixed model prediction and flagging:

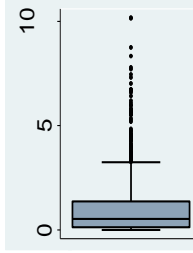
<https://github.com/cemcculloch/Flagging-unusual-clusters>

Linear mixed model self-calibrated predictors including analysis of the walking speed data:

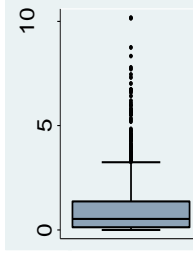
<https://github.com/cemcculloch/Self-calibrated-predictors>

Supplemental files for Biometrics 2025 article including Stata code for the asthma example (binary outcome model):

https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/biometrics/81/3/10.1093_biomtc_ujaf094/1/ujaf094_supplemental_files.zip?Expires=1768323431&Signature=wMCt1rSISrPLiX42kZxjn4hL1gpO31TXXDV-BdLq8z1xVmezjRjg8kBvOGOaHZers5dg5DOif~M9xPyTflz~ve1nfx-yyqpy6x8nWIAy90ATaCXvUMYDSOwhmgdGNGsZAX0dHrZX3ckctA3kYTOw4~RMslvx0sP-4KvKX34enNxRSWaL499RHUF CSPDMWSMI9IniA9X9xYWj66zLIBvsITS8stMFMLyiZclcJsLR2qjPas4RKulnMHDom9Tgs1mtaChE6TZ2dB~zE3znDEsLCQXZr16m~1cMHUr5CFsL248TecUEagKi1kSce9r3G8dS Wj0iK6TNCi~vIEPKXiMqKA__&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA



References



Neuhaus J, McCulloch C, Boylan R. Improved prediction and flagging of extreme random effects for non-Gaussian outcomes using weighted methods. *Biometrics*. 2025 Jul 3;81(3):ujaf094. doi: 10.1093/biomtc/ujaf094. PMID: 40736765; PMCID: PMC12309285.

McCulloch CE, Neuhaus JM, Boylan RD. Flagging unusual clusters based on linear mixed models using weighted and self-calibrated predictors. *Biometrics*. 2024 Mar 27;80(2):ujae022. doi: 10.1093/biomtc/ujae022. PMID: 38563530; PMCID: PMC12633726.

McCulloch, C. E., & Neuhaus, J. M. (2021). Improving Predictions When Interest Focuses on Extreme Random Effects. *Journal of the American Statistical Association*, 118(541), 504–513. <https://doi.org/10.1080/01621459.2021.1938583>.