Statistical research at CBD

At the interface between biology and mathematical sciences

Stefanie Muff

February 16 2022

(https://quotefancy.com/)

CBD is a real paradise for a statistician like me.

Research topics

- Quantitative genetics/evolutionary biology
- Movement ecology/habitat selection studies
- Interpretation and use of statistical methods
  - p-values
  - Variable importance in regression models
- Measurement error modeling
- Bayesian statistics

Research topics

- Quantitative genetics/evolutionary biology
- Movement ecology/habitat selection studies
- Interpretation and use of statistical methods
  - p-values
  - Variable importance in regression models
- Measurement error modeling
- Bayesian statistics

$\rightarrow$ *All these topics are linked with each other.* How?

Example 1: Heritability is a measurement error problem

- A phenotype (P) can be additively decomposed into a genetic (G) and an environmental (E) component: $P = G + E$.
- The phenotypic variance σ_p^2 can be decomposed accordingly: $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$.
- The aim is to estimate σ_g^2, and in particular *heritability*

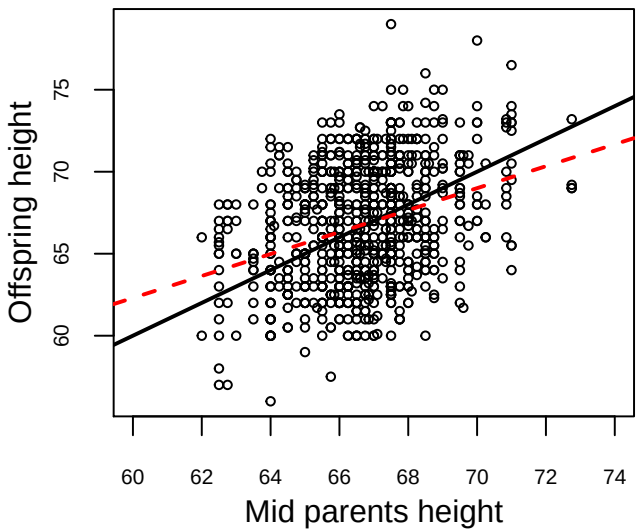
$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \ .$$

Example 1: Heritability is a measurement error problem

- A phenotype (P) can be additively decomposed into a genetic (G) and an environmental (E) component: $P = G + E$.
- The phenotypic variance σ_p^2 can be decomposed accordingly: $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$.
- The aim is to estimate σ_g^2, and in particular *heritability*

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \ .$$

- We can estimate h^2 by *mid-parent regression*:



- In the measurement error world, assume we observe
  $w_i = x_i + u_i$, with $u_i \sim \mathsf{N}(0, \sigma_u^2)$, where x_i is the true variable.

- The aim is to find the slope β_x

$$y_i = \beta_0 + \beta_x x_i + \varepsilon_i \ ,$$

  but *in practice* we regress

$$y_i = \beta_0 + \beta_x^\star w_i + \varepsilon_i \ .$$

- In the measurement error world, assume we observe
  $w_i = x_i + u_i \ , \quad$ with $u_i \sim \mathsf{N}(0, \sigma_u^2)$, where x_i is the true variable.

- The aim is to find the slope β_x

$$y_i = \beta_0 + \beta_x x_i + \varepsilon_i \ ,$$

  but *in practice* we regress

$$y_i = \beta_0 + \beta_x^\star w_i + \varepsilon_i \ .$$

- Then it is known that we are actually estimating

$$\beta_x^\star = \underbrace{\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}}_{=\lambda} \cdot \beta_x \ .$$

- In the measurement error world, assume we observe
  $w_i = x_i + u_i \ , \quad$ with $u_i \sim \mathsf{N}(0, \sigma_u^2)$, where x_i is the true variable.

- The aim is to find the slope β_x

$$y_i = \beta_0 + \beta_x x_i + \varepsilon_i \ ,$$

  but *in practice* we regress

$$y_i = \beta_0 + \beta_x^\star w_i + \varepsilon_i \ .$$

- Then it is known that we are actually estimating

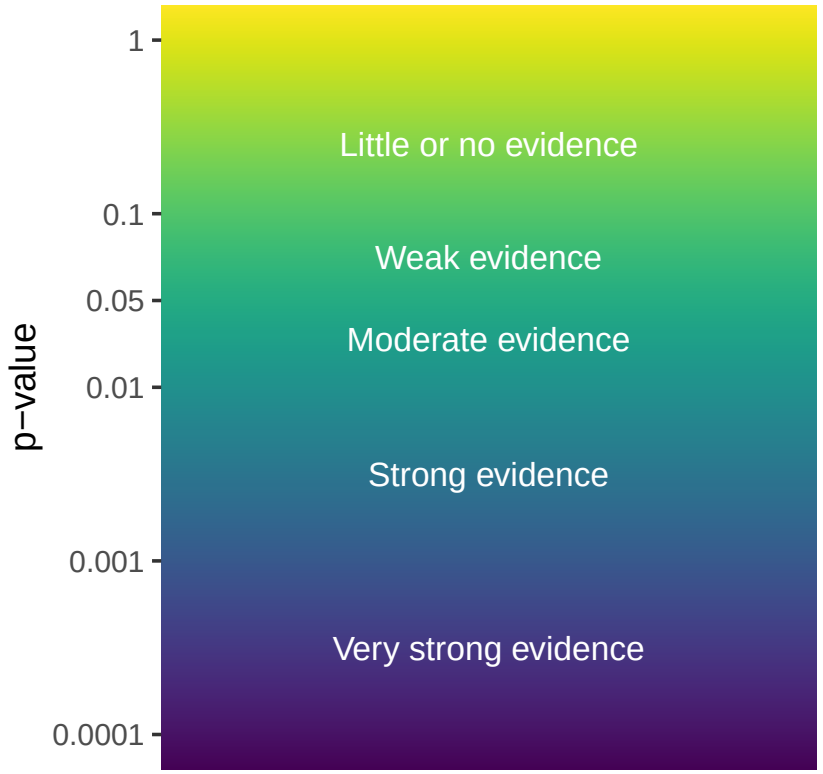
$$\beta_x^\star = \underbrace{\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}}_{=\lambda} \cdot \beta_x \ .$$

- For a slope of $\beta_x = 1$, the attenuation factor $\lambda = h^2$.

Example 2: The interpretation of p-values is relevant in most studies

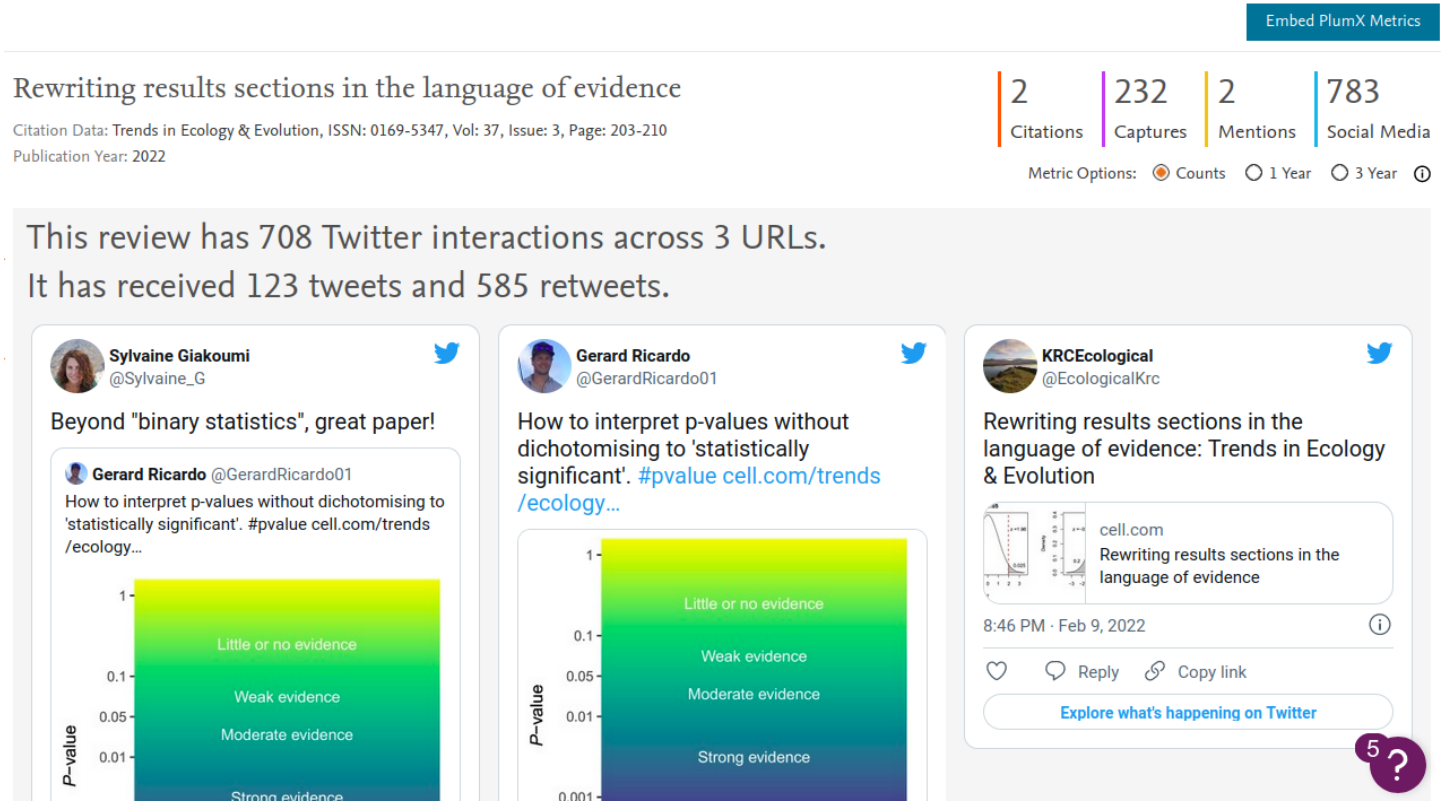
Opinion

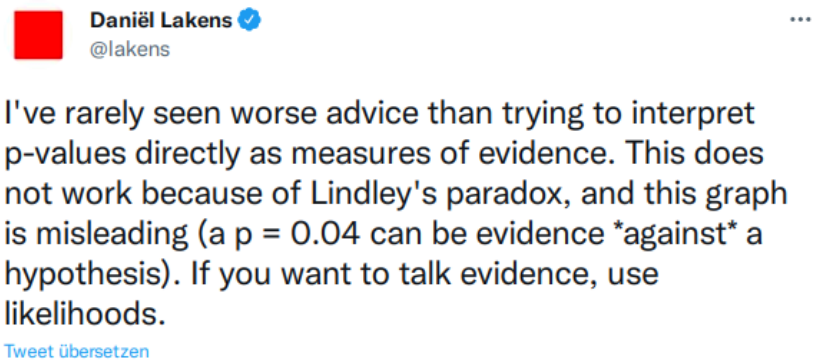
Rewriting results sections in the language of evidence



Muff et al. (2022)

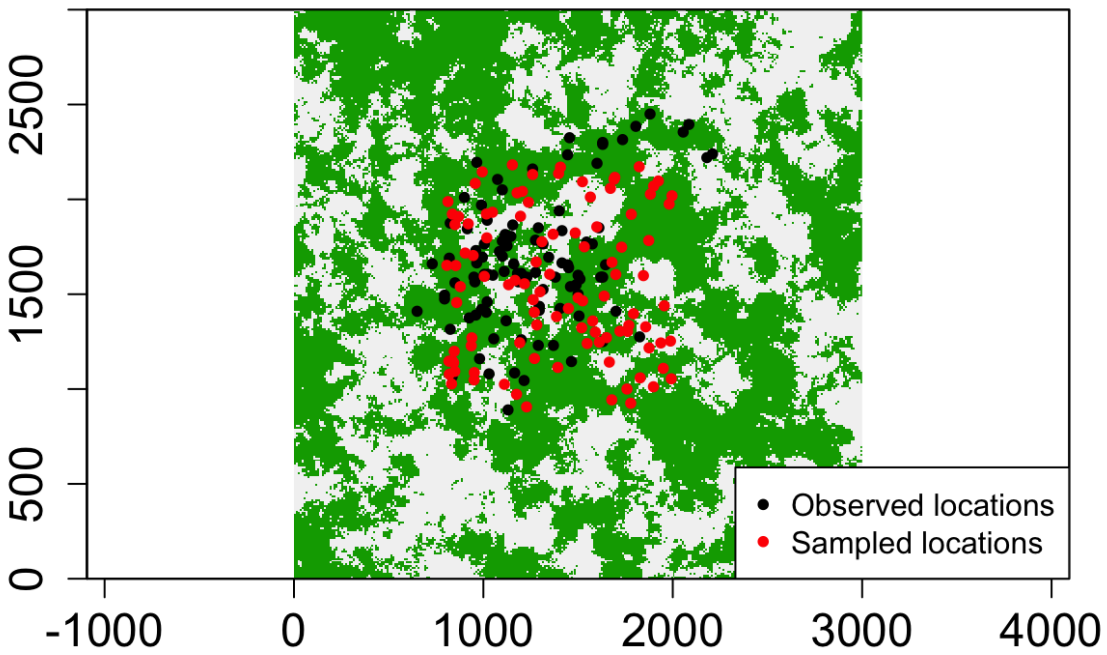
Huge echo! We received three letters, people wrote three blog posts and the Twitter community had very controversial opinions.





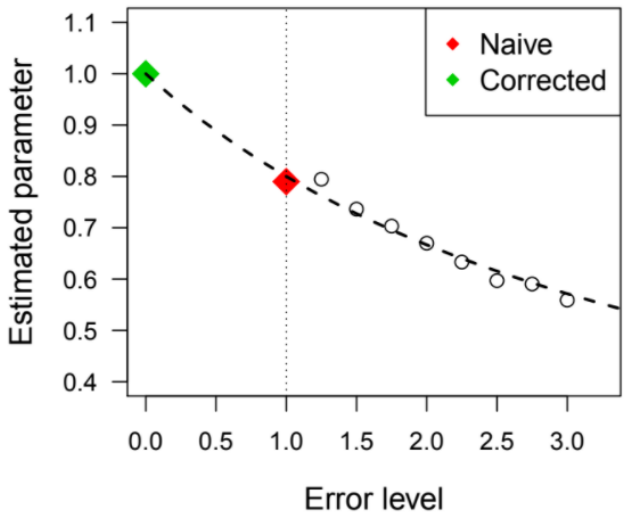
Example 3: Movement ecology

Aim: Accounting for GPS error in movement ecology studies.



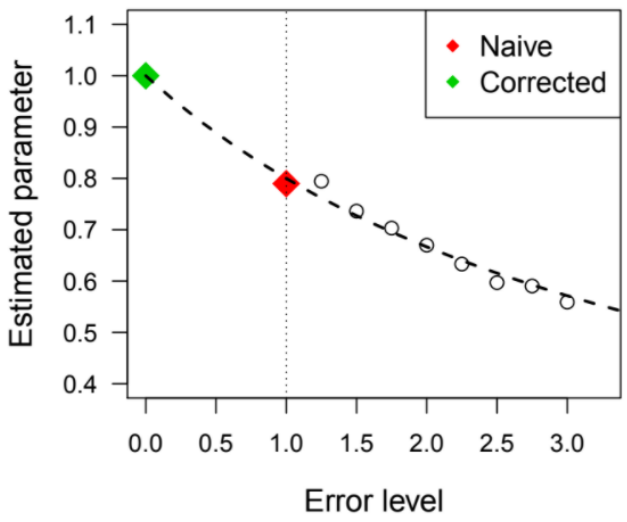
Complex problem, because GPS error propagates into covariate error in the regression of interest.

- Idea: Simulation Extrapolation (SIMEX):

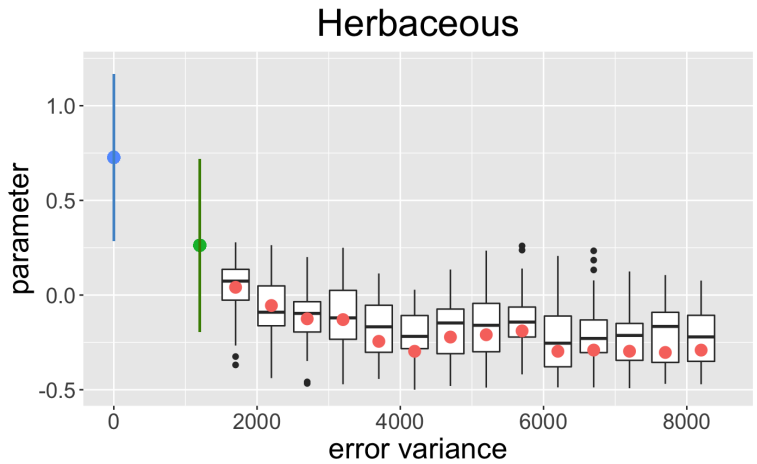


[1] Master thesis by Clara Panchaud

- Idea: Simulation Extrapolation (SIMEX):



- Generalize this to GPS error.[1]



[1] Master thesis by Clara Panchaud

Back to quantitative genetics

- Remember the $P = G + E$ decomposition
- In statistical terms:

$$y_i = \mu + g_i + \varepsilon_i \ ,$$

where

$$(g_1, \ldots, g_n)^\top \sim \mathsf{N}(\mathbf{0}, \sigma_g^2 \cdot \mathbf{G}) \ ,$$
$$\varepsilon_i \sim \mathsf{N}(\mathbf{0}, \sigma_e^2) \ .$$

Back to quantitative genetics

- Remember the $P = G + E$ decomposition
- In statistical terms:

$$y_i = \mu + g_i + \varepsilon_i \ ,$$

where

$$(g_1, \ldots, g_n)^\top \sim \mathsf{N}(\mathbf{0}, \sigma_g^2 \cdot \mathbf{G}) \ ,$$
$$\varepsilon_i \sim \mathsf{N}(\mathbf{0}, \sigma_e^2) \ .$$

"Animal model"

Back to quantitative genetics

- Remember the $P = G + E$ decomposition
- In statistical terms:

$$y_i = \mu + g_i + \varepsilon_i \ ,$$

where

$$(g_1, \ldots, g_n)^\top \sim \mathsf{N}(\mathbf{0}, \sigma_g^2 \cdot \mathbf{G}) \ ,$$
$$\varepsilon_i \sim \mathsf{N}(\mathbf{0}, \sigma_e^2) \ .$$

"Animal model"

- In wild system, information about the relatedness among individuals ($\mathbf{G}$) is historically derived from pedigrees.

Quantitative genetics with genomic data

- Relatedness information is more and more derived from genomic data $\rightarrow$ A paradise for statisticians!
- Problem: The $n \times n$ relatedness matrix $\mathbf{G}$ is dense. Computation becomes prohibitive for many individuals.

[2] PhD project by Janne Hetle
[3] Meuwissen et al. (2001)

Quantitative genetics with genomic data

- Relatedness information is more and more derived from genomic data $\rightarrow$ A paradise for statisticians!
- Problem: The $n \times n$ relatedness matrix $\mathbf{G}$ is dense. Computation becomes prohibitive for many individuals.
- Idea: *Marker-based regression*[2] (equivalent to animal model[3]):

$$y_i = \mu + \underbrace{\sum_j \underbrace{(SNP_{ij} \cdot u_j)}_{g_{ij}}}_{g_i} + \varepsilon_i \ .$$

[2] PhD project by Janne Hetle
[3] Meuwissen et al. (2001)

Quantitative genetics with genomic data

- Relatedness information is more and more derived from genomic data $\rightarrow$ A paradise for statisticians!
- Problem: The $n \times n$ relatedness matrix $\mathbf{G}$ is dense. Computation becomes prohibitive for many individuals.
- Idea: *Marker-based regression*[2] (equivalent to animal model[3]):

$$y_i = \mu + \underbrace{\sum_j \underbrace{(SNP_{ij} \cdot u_j)}_{g_{ij}}}_{g_i} + \varepsilon_i \ .$$

- Problem: Too many unknowns, $p >> n$.

[2] PhD project by Janne Hetle
[3] Meuwissen et al. (2001)

We need statistical methods that make us ready for the upcoming challenges with genomic data from wild study systems!

[4]Ødegaard et al. (2018)

We need statistical methods that make us ready for the upcoming challenges with genomic data from wild study systems!

Ideas (Jannes PhD):

- Use singular value decomposition (SVD) for dimension reduction[4].
- Use a Bayesian approach to fit a full model with additional fixed and random effects (crucial in wild systems).
- Use a ridge-based shrinkage prior to ensure efficiency.

[4] Ødegaard et al. (2018)

We need statistical methods that make us ready for the upcoming challenges with genomic data from wild study systems!

Ideas (Jannes PhD):

- Use singular value decomposition (SVD) for dimension reduction[4].
- Use a Bayesian approach to fit a full model with additional fixed and random effects (crucial in wild systems).
- Use a ridge-based shrinkage prior to ensure efficiency.

→ *In brief, we are combining several statistical concepts (SVD, Bayes, shrinkage, mixed modeling,. . .) to hopefully develop an efficient and useful method for evolutionary biology.*

[4] Ødegaard et al. (2018)

Summary and Outlook

- Almost unlimited possibilities for statisticians at CBD.
- Several Master and PhD students with exciting projects (Clara Panchaud, Vebjørn Rekkebo, Kenneth Aase, Janne Hetle, Emma Skarstein...)
- More ideas than resources.
- (Too?) many collaborations.

A final impression

An email from a CBD postdoc last week (10.2.22):

> *I am a biologist with very superficial training in math and stats, but I like to dabble in making models and thinking analytically. I always feel like I am reinventing the wheel when I do... Or rather I am making a terrible square wheel that doesn't quite do the job as well as it could.*
> *Which is not a very good feeling!*
> *So my question is this: what can we do about it? Obviously people with your skills are in short supply in biology departments.*
> *I don't know if there would be a way to organize little discussion sessions where you and perhaps other experts can give feedback to people like me that would present their ideas.*
> *That could spur new collaborations, or at least spare us some embarrassment.*