RSS discussion meeting, March 24, 2026

Discussion of 'Regression by composition' by Farewell et al.

Heather Battey
Department of Mathematics, Imperial College London

McCullagh's (1999, 2002, 2022) foundations of modelling

Treatment has an effect on probability distribution of outcome.

Transformation induced by action of group on baseline distributions $\{P_\gamma : \gamma \in \Gamma\}$.

Simplest models: group $\mathcal{G}$ acts on the baseline parameter space Γ

$$(g, \gamma) \to g\gamma \in \Gamma, \quad \gamma \in \Gamma, \ \ g \in \mathcal{G}.$$

A group is a set with a binary operation.

McCullagh's (1999, 2002, 2022) foundations of modelling

Treatment has an effect on probability distribution of outcome.

Transformation induced by action of group on baseline distributions $\{P_\gamma : \gamma \in \Gamma\}$.

Simplest models: group $\mathcal{G}$ acts on the baseline parameter space Γ

$$(g, \gamma) \to g\gamma \in \Gamma, \quad \gamma \in \Gamma, \ \ g \in \mathcal{G}.$$

Regression: γ_i depends on baseline covariates, e.g. $\gamma_i = w_i^{\mathrm{T}}\beta$ or $\gamma_i = \exp(w_i^{\mathrm{T}}\beta)$.
Group actions: addition and multiplication respectively.

A group is a set with a binary operation.

Regression by composition

Extension: effects of all variables treated as group elements.

Possibly different group action for each effect.

Ordering

An ordering is inherent. Natural for successive treatment/exposure variables and intermediate outcomes but unnatural for intrinsic variables fixed at baseline.

Q1: *Advice? Treat baseline variables symmetrically?*

Q2: *Does advice change depending on objective: understanding vs prediction?*

Equivalence classes of models

Two contexts:

1. Intrinsic variables treated symmetrically; other effects modelled by RBC.
2. Inherent ordering of RBC not interpreted literally.

Q3.1: *Equivalence classes of models generating same distribution over Y given X?*

Q3.2: *RBC models that are statistically indistinguishable at a given sample size and significance level?*

Q3.3: *How is model adequacy to be assessed?*

Instability and physical constraints

- Most collapsible models not physical, e.g. probabilities not constrained to $[0, 1]$.
- Enforce physical constraints $\implies$ parameter space depends on sample.
  Undermines stable interpretation of parameters; violates McC's (2002) axioms.
- Stability preserved by allowing fitted values to fall out of range.
  Acknowledges that model is a poor approximation for some individuals.
  Not permissible with maximum-likelihood fitting.

Q4: *Implications for regression by composition?*

Interpretation in collapsible models

Regression coefficient β_j in linear in probability model (collapsible).

- If interpreted purely mathematically: *expected change in the proportion of m individuals likely to experience the positive outcome in response to a hypothetical replacement of m individuals who differ by one unit in the jth component and are otherwise identical; m arbitrary.*

Interpretation in collapsible models

Regression coefficient β_j in linear in probability model (collapsible).

- If interpreted purely mathematically: *expected change in the proportion of m individuals likely to experience the positive outcome in response to a hypothetical replacement of m individuals who differ by one unit in the jth component and are otherwise identical; m arbitrary.*
- Interpretation shifts from individual-level to population-level: variables that were intrinsic at the individual level are not so when treated as population averages.

Q5: *How do we feel about this? What manifestations, if any, does it have in the collapsible components of RBC?*

Incidental: the physically implausible linear in probability model is the only one for which the treatment parameter agrees with the Neyman-Rubin treatment effect under the relevant fictitious idealisation.

Q6: Why care about collapsibility?

Q6.1: *If covariates thought to be relevant are measured, should they not be included?*

Q6.2: *A case for deliberately omitting them would be if they were too numerous. Points back to model assessment?*

Collapsibility and unmeasured covariates

Suppose that relevant baseline covariates are not measured.

Cox (1958): binary matched pairs. Logistic model (not collapsible).
All-encompassing nuisance parameter γ_i for ith pair captures unmeasured aspects.

$$q_{1|0}^{(i)} := \text{pr(success} \mid \text{control)} = \frac{e^{\gamma_i}}{1+e^{\gamma_i}}, \qquad q_{0|0}^{(i)} = 1 - q_{1|0}^{(i)}$$

$$q_{1|1}^{(i)} := \text{pr(success} \mid \text{treated)} = \frac{e^{\gamma_i+\beta}}{1+e^{\gamma_i+\beta}}, \quad q_{0|1}^{(i)} = 1 - q_{1|1}^{(i)}, \quad i = 1, \ldots n.$$

- Direct maximum lik over $n+1$ parameters gives erroneous inference ($\hat{\beta} \to_p 2\beta$).
- Collapsed analysis leads to erroneous inference on β.
- Condition twice: eliminates $(\gamma_i)_{i=1}^n$; conditional MLE consistent for β.

Congratulatory remarks

An original, scholarly, and thought-provoking paper.

References

- Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.
- McCullagh, P. (1999). The algebraic structure of generalised linear models. University of Chicago, report number 489.
- McCullagh, P. (2002). What is a statistical model? *Ann. Statist.*, 30, 1225–1310.
- McCullagh, P. (2022). *Ten Projects in Applied Statistics*. Springer, Cham.