# Regression by composition

Daniel M. Farewell

*Division of Population Medicine, School of Medicine, Cardiff University, UK*

E-mail: farewelld@cardiff.ac.uk

Rhian M. Daniel

*Division of Population Medicine, School of Medicine, Cardiff University, UK*

Anders Huitfeldt

*Division of Mental Health and Addiction, Møre og Romsdal Hospital Trust, Norway*

**Summary**. Regression models describe probability distributions conditional on covariates. By shifting or scaling, for example, the conditional distribution at one value of the covariates may be transformed into the conditional distribution at another covariate value. Flexibly combining several such transformations leads to a family of models we call regression by composition. The regression by composition class includes many old friends, and some new ones, too. Studying the algebraic properties of the component transformations leads to insights about collapsibility, closure, and model constraints. We provide an indicative typology of transformations suitable for discrete, continuous and time-to-event outcomes, and apply these ideas to a historical analysis of arable land rent prices, and to a study of synbiotics for sepsis prevention.

## 1. Introduction

Regression models are convenient, general-purpose tools of the statistical trade. Indeed, there is so much flexibility in the idea of examining variation in one random variable through variation in others that it can be difficult to articulate what—if anything—unites this class of models. Equally, what divides them? What important model features are possessed by some but not all? And, given a particular model, how shall we determine if it possesses these features?

The present paper offers one approach to framing, and answering, such questions. We construct regression models through a sequence of random probability laws, with covariate-dependent transformations turning one law into another. It turns out that these transformations hold the keys to many statistical features of regression models, including closure (valid probability laws always map to other valid probability laws), collapsibility (marginal quantities are averages of conditional quantities), and the presence or absence of constraints (fixed points of the model). In applications, lack of closure can yield meaningless out-of-sample extrapolation, while non-collapsible treatment comparisons rule out simple weighted-average approaches to meta-analysis, and constraints highlight areas of low model flexibility. Explicit transformation of laws also enables transparent discussion of the plausibility of parametric assumptions we may choose to make; in short, it encourages informed model criticism (Box, 1980).

The proposed unifying framework, which we call regression by composition, includes as special cases all generalized linear and generalized additive models (Nelder & Wedderburn, 1972; Hastie & Tibshirani, 1986), multivariate and repeated measures regression (Laird & Ware, 1982), additive (Aalen, 1989) and relative (Cox, 1972) hazard models for time-to-event data, accelerated

failure time models (see Wei, 1992, and references therein), transformation models (Fine et al., 1998; Zeng & Lin, 2007), and joint mean-covariance models (Nelder & Lee, 1991; Nelder et al., 1998). One model can often be turned into another through a simple interchange of one or two modular elements but, at least in principle, the same simple fitting algorithm can be used in every case.

Our approach can also construct models that do not fall within any of the aforementioned classes. This includes models in which each covariate is associated with its own family of transformations (enabling each covariate to have its own local link function), and also models closely related to the switch relative risk (van der Laan et al., 2007), a kind of binary regression that can be motivated from first principles and has desirable properties when computing marginal causal effects. Although our focus in this paper is not explicitly on causal inference, our preference is to relate model components to real-world processes wherever possible.

## 2. Example: the Cox–Aalen model

We begin by describing an existing regression model, in so doing exhibiting several important features of a general regression by composition. Scheike and Zhang (2002) introduced the so-called *additive-multiplicative* or *Cox–Aalen* model for censored survival times, which features a flexible additive (Aalen) component to describe a potentially covariate-dependent baseline hazard, and a 'proportional hazards' (Cox) component that combines multiplicatively with this baseline hazard. The additive-multiplicative model extends the better-known stratified Cox model (Kalbfleisch & Prentice, 2002, p. 118) to allow the baseline hazard to depend *parametrically* on covariates.

To fix ideas, imagine a clinical trial in which patients at varying risk of premature death are randomized to receive either usual care or a novel therapy designed to extend their lives. For the purposes of exposition, we shall suppose that patients' survival times from study start ($t = 0$) to time of death $Y > 0$ are observed without censoring of any kind. Patient age $A$ (a positive real number) and treatment group $G \in \{0, 1\}$ are recorded at $t = 0$, with $G = 1$ indicating receipt of the novel therapy.

An additive-multiplicative model specifies the hazard increment $\mathrm{d}\Lambda(t) = \lambda(t)\,\mathrm{d}t$ at time $t > 0$ (that is, the instantaneous probability of death in a small interval after $t$, given survival to $t$) as

$$\mathrm{d}\Lambda(t) = \{\mathrm{d}\eta_1(t) + \mathrm{d}\eta_2(t)\} \times \eta_3(t),$$

where $\eta_1, \eta_2$ and $\eta_3$ are functions of time, and random through their dependence on covariates (here $A$ and $G$). The hazard increment for individuals in the usual care (or 'control', $G = 0$) group and aged at the chosen reference level for this covariate ($A = 64$, say) is $\mathrm{d}\Lambda(t) = \mathrm{d}\eta_1(t)$. Individuals of known age $A$ who are in the control group will gain the extra hazard increment component $\mathrm{d}\eta_2(t)$, the precise value of which depends on $A$, and so has overall hazard increment $\mathrm{d}\Lambda(t) = \mathrm{d}\eta_1(t) + \mathrm{d}\eta_2(t)$. Individuals of known age $A$ and known treatment group $G$ gain the multiplicative component $\eta_3(t)$, which depends on $G$ through $\eta_3$, and so have overall hazard increment $\mathrm{d}\Lambda(t) = \{\mathrm{d}\eta_1(t) + \mathrm{d}\eta_2(t)\} \times \eta_3(t)$.

Regression by composition mirrors this sequential construction of the additive-multiplicative model. Given a distribution $\mathbb{P}_1$ describing the survival experience of an individual in the control group and at the reference value for age, we ask how $\mathbb{P}_1$ would need to be changed in order to now reflect the survival experience of an individual in the control group but aged $A$ years old. This is a central idea for all that follows: transformations whose inputs and outputs are both distributions

are gathered into families—called flows—that are indexed by parametrized linear combinations of covariates. In the case of survival analysis, it is convenient to characterize distributions in terms of the survivor function, but other choices are possible, and lead to equivalent models.

The distribution $\mathbb{P}_1$ of reference-aged individuals in the control group has associated survivor function $\hat{\mathbb{P}}_1 : t \mapsto \mathbb{P}_1(Y > t)$, which in this case is $\hat{\mathbb{P}}_1 : t \mapsto \exp\{-\eta_1(t)\}$. Here we use a 'hat' not to denote estimation, but a convenient 'representation' of a probability distribution; characteristic functions are alternatives to survivor functions, and motivate the 'hat' notation. The additive component of the Cox–Aalen model now updates $\hat{\mathbb{P}}_1$ via the flow $f_2$ and produces the survivor function $\hat{\mathbb{P}}_2$ describing the survival experience of an untreated individual whose age $A$ is known, given by

$$\hat{\mathbb{P}}_2(t) = f_2(\hat{\mathbb{P}}_1, \eta_2)(t) = \hat{\mathbb{P}}_1(t) \exp\{-\eta_2(t)\} = \exp[-\{\eta_1(t) + \eta_2(t)\}].$$

The survivor function $\hat{\mathbb{P}}_2$ corresponds to the cumulative hazard function $\eta_1 + \eta_2$. The flow $f_2$ takes two arguments: principally, the input distribution $\mathbb{P}_1$ (or, as here, survivor function $\hat{\mathbb{P}}_1$), but also an index $\eta_2$ (historically called an 'effect size'). In this case $\eta_2$ is a function. Notice that if $\eta_2(t) = 0$ for all $t$ (so $\eta_2$ is the zero *function*) then the corresponding transformation $f_2(\cdot, t \mapsto 0)$ is the identity function and maps the survivor function $\hat{\mathbb{P}}_1$ to itself, so that $\hat{\mathbb{P}}_2 = \hat{\mathbb{P}}_1$. Though we defer the details, we may construct the *local linear predictor* $\eta_2$ in such a way that $\eta_2 = 0$ for individuals in the reference category of age. As the age covariate $A$ increases and $\eta_2$ moves away from the zero function (perhaps $\eta_2$ is an increasing function, with $\eta(0) = 0$), hazards are increased additively and the distribution of $Y$ is correspondingly deformed towards shorter survival times.

We now incorporate the information about treatment group, and modify the distribution $\mathbb{P}_2$ by way of the multiplicative flow $f_3$ to yield the distribution $\mathbb{P}_3$ describing an arbitrary individual of any age and in either group, given by

$$\hat{\mathbb{P}}_3(t) = f_3(\hat{\mathbb{P}}_2, \eta_3)(t) = \prod_{s=0}^{t} \{\mathrm{d}\hat{\mathbb{P}}_2(s)\}^{\eta_3(s)} = \exp[-\int_{s=0}^{t} \{\mathrm{d}\eta_1(s) + \mathrm{d}\eta_2(s)\} \times \eta_3(s)].$$

Within the product integral, $\mathrm{d}\hat{\mathbb{P}}_2(s)$ is the *multiplicative* increment $\hat{\mathbb{P}}_2(s + \mathrm{d}s)/\hat{\mathbb{P}}_2(s)$, being directly analogous to the survival proportions appearing in a Kaplan–Meier curve (Kaplan & Meier, 1958). Here again the flow $f_3$ takes two arguments: the input distribution $\mathbb{P}_2$ and a positive-valued index function $\eta_3$. Like $f_2$, the flow $f_3$ is a deterministic function, but again its evaluation is random through its dependence on the random function $\eta_3$, which in turn is random through its dependence on treatment information $G$. For individuals with $G = 0$ (the reference value for treatment), our setup can enforce $\eta_3(t) = 1$ (the multiplicative identity), and $f_3(\cdot, t \mapsto 1)$ would again be the identity map. If instead $\eta_3$ is a positive constant (as indeed is often assumed) then the Cox flow simplifies to

$$f_3(\hat{\mathbb{P}}_2, \eta_3)(t) = \{\hat{\mathbb{P}}_2(t)\}^{\eta_3}.$$

In general the survivor function $\hat{\mathbb{P}}_3$ corresponds to the cumulative hazard function $t \mapsto \int_0^t \{\mathrm{d}\eta_1(s) + \mathrm{d}\eta_2(s)\} \times \eta_3(s)$, or more simply $\mathrm{d}\Lambda(t) = \{\mathrm{d}\eta_1(t) + \mathrm{d}\eta_2(t)\} \times \eta_3(t)$, reproducing the Cox–Aalen model as desired.

The additive component of the Cox–Aalen model is notable for several reasons. First, there are potential problems with using and understanding the model if ever $\mathrm{d}\eta_2(t) < 0$, and most especially if $\mathrm{d}\eta_2(t) < -\mathrm{d}\eta_1(t)$, for this would correspond to a negative overall hazard increment. Aalen (1989) points out that this infelicity may or may not be of practical importance, but also

acknowledges that admissible values of $\eta_2$ leading to corresponding $\mathbb{P}_2$ that are not valid probability measures nevertheless makes the model feel somewhat less 'natural'. Second, and more positively, the model is *linear* in two senses: most obviously the hazards combine linearly, but also the flow $f_2$ is linear in its input survivor function $\hat{\mathbb{P}}_1$. It turns out that this latter linearity of the flow is what gives the Aalen model one of its attractive features: if the model is correct, it is *collapsible* (Martinussen & Vansteelandt, 2013). Roughly, this means that between-group comparisons (of the particular kind implied by the model) and within-group averages can be made interchangeably.

Now let us consider the multiplicative component of the model. Since $\eta_3$ is assumed everywhere positive, there is no danger that a valid survivor function may become invalid following application of the Cox flow $f_3$: in other words, the flow $f_3$ is *closed* on the set of probability distributions. On the other hand, in recent years several authors have drawn attention to the fact that the Cox model is *not* collapsible (Hernán, 2010; Martinussen & Vansteelandt, 2013; Aalen et al., 2015). This non-collapsibility arises because the multiplicative flow $f_3$ is not an affine (roughly, straight line) function of its input $\hat{\mathbb{P}}_2$: notice the appearance of $\eta_3$ as an *exponent* in $f_3$.

Our informal construction of the additive-multiplicative model is designed to illustrate how algebraic properties of the flows $f_2$ and $f_3$ (e.g. closed, affine) correspond directly to important local features of the statistical model. Other possible questions with algebraic parallels include 'is the Cox–Aalen model the same thing as the Aalen–Cox model?', or equivalently 'do the flows $f_2$ and $f_3$ commute?'. The answer is 'no, they are not the same, because no, they do not commute': the order in which the two flows appear matters. Another interesting contrast between the two models concerns their so-called *zero constraints* (that is, where the model "constrain[s] the benefits of treatment to be zero"; see Deeks, 2002): the Aalen flow has one fixed point (where everyone dies instantly; $f_2(t \mapsto 0, \eta) = t \mapsto 0$ for all $\eta$), while the Cox flow has two (where everyone dies instantly, or where everyone lives forever; $f_3(t \mapsto 0, \eta) = t \mapsto 0$ and $f_3(t \mapsto 1, \eta) = t \mapsto 1$ for all $\eta$).

The Cox-Aalen model is a very special case of our much more general framework. But all regressions by composition are built up by composing together different flows in exactly this way. There is considerable flexibility in the choice of flow corresponding to each covariate or set of covariates, and often there are tradeoffs to be made between properties like closure and collapsbility. This is one reason among many that we find it so helpful to articulate which flows have this property or that property, and to establish why. To this end, we now turn to a description of regression by composition in a much more general setting.

## 3. Regression by composition

### *3.1. Scope and setting*

We employ the term *regression model* to describe any mathematical formulation of the conditional distribution of an *outcome* or *response* random variable $Y$ given *covariates*. In what follows, the outcome $Y$ can take any form, and could for instance be a tuple, an image, or a stochastic process. However, to emphasize the intuition of what we call regression by composition, our examples will focus almost exclusively on scalar response variables. Nevertheless, the proofs and procedures we offer are not limited to the scalar case.

We assume that outcome and covariates are all random variables defined on a probability space $(\Omega, \mathcal{E}, \mathbb{E})$. Conceptually, the probability measure $\mathbb{E}$ is the 'true' probability measure underlying our observations, aspects of which we are interested in making inference about. We write $\mathbb{E}$ for

both the probability measure and for expectation with respect to the measure (de Finetti notation; see Pollard, 2002, pp. 7–11). The outcome $Y$ takes values in a measurable space $(\Upsilon, \mathcal{Y})$, and the covariates generate a sigma-algebra $\mathcal{F} \subset \mathcal{E}$. We will typically assume that $\Upsilon$ is (or is part of) an inner product space so that $Y$ has a well-defined characteristic function or other generating function, but the covariates need have no special structure.

In common with most regression settings, we make no attempt to model the stochastic behaviour of covariates, and instead focus on the conditional law of the random variable $Y$ given covariate information $\mathcal{F}$. This decision is not restrictive: any joint distribution can be constructed as a product of such conditional distributions, and indeed this sequential point of view is emphasized by Bayesian networks (Pearl, 1985) and causal directed acyclic graphs (see, for example, Pearl, 2009). We shall be concerned with modelling the ($\mathcal{F}$-measurable) random, conditional law $\mathbb{P}\colon \mathcal{Y} \times \Omega \to [0,1]$ given by

$$\mathbb{P}(A, \omega) = \mathbb{E}(Y \in A \mid \mathcal{F})(\omega).$$

By a regression model, we mean precisely a set of candidates for the conditional law $\mathbb{P}$. We treat any such candidate as a general (random) map from the set of events $\mathcal{Y}$ to the unit interval, rather than as a (say) normal distribution characterized by its (random) mean and variance. Ultimately, our set of candidates for $\mathbb{P}$ will be described by finite-dimensional parameters; however, *any* probability distribution could in theory arise as part of a regression by composition.

Studying random laws such as $\mathbb{P}$ is simplified if we situate them within a suitable vector space $\mathcal{M}$ of measures on $(\Upsilon, \mathcal{Y})$. To this end, it is natural to hope that, for some subset of covariates generating a smaller sigma-algebra $\mathcal{G} \subseteq \mathcal{F}$, we should obtain

$$\mathbb{E}(\mathbb{P} \mid \mathcal{G})(A) = \mathbb{E}(Y \in A \mid \mathcal{G})$$

for each $A \in \mathcal{Y}$ when averaging (or 'marginalizing') the random law $\mathbb{P}$ over this coarser information. The choice of vector space $\mathcal{M}$ is actually what *defines* what is meant by the (conditional) mean $\mathbb{E}(\mathbb{P} \mid \mathcal{G})$ of a random law $\mathbb{P}$: expectation is just integration, which in turn depends principally on notions of addition $P + Q$ and scalar multiplication $aP$ of measures $P, Q \in \mathcal{M}$. By the tower law,

$$\mathbb{E}(Y \in A \mid \mathcal{G}) = \mathbb{E}\{\mathbb{E}(Y \in A \mid \mathcal{F}) \mid \mathcal{G}\} = \mathbb{E}\{\mathbb{P}(A) \mid \mathcal{G}\},$$

so to achieve our aim we must identify $\mathbb{E}(\mathbb{P} \mid \mathcal{G})(A)$ with $\mathbb{E}\{\mathbb{P}(A) \mid \mathcal{G}\}$. In words, addition and scalar multiplication of measures within the vector space $\mathcal{M}$ should mirror the corresponding addition and scalar multiplication of their *probabilities*. This leads us directly to the vector space $\mathcal{M}$ of (signed, finite) measures on $(\Upsilon, \mathcal{Y})$, with addition of measures $P$ and $Q$ and scalar multiplication by reals $a$ and $b$ defined via $(aP+bQ)(A) = aP(A)+bQ(A)$, for $A \in \mathcal{Y}$. Endowing the vector space $\mathcal{M}$ with the total variation norm turns $\mathcal{M}$ into a Banach space (Kreyszig, 1978, p. 58) and ensures that integration of functions (specifically, expectations of random variables, which in this case happen to be random laws) taking values in $\mathcal{M}$ is well-defined in the sense of Bochner (1933). We shall also attach special significance to the simplex $\mathcal{P} \subset \mathcal{M}$ of probability measures, although the latter is not a vector subspace of $\mathcal{M}$; among other reasons, the zero measure is not in $\mathcal{P}$.

### *3.2. Model Specification*

#### 3.2.1. *Flows*

Flows are function families, and the building blocks of regression by composition. Their components are transformations of $\mathcal{M}$, the space of laws on $(\Upsilon, \mathcal{Y})$. Informally, the transformations in

a flow share a common 'shape', while its index parametrizes their 'size'. To motivate the general definition, we first feature a familiar workhorse, reshod for regression by composition: the *location shift* flow.

If $P$ is a probability measure on $\Upsilon = \mathbb{R}$, then so is any measure $f^v(P)$ defined by transforming its cumulative distribution function as $f^v(P)(Y \leq t) = P(Y \leq t - v)$. The transformation $f^v$ simply shifts the whole distribution upwards by a quantity $v$. Any $v \in \mathbb{R}$ would define such a transformation of measures, and collectively the transformations $\{f^v : v \in \mathbb{R}\}$ make up the location shift flow. The two defining features of a flow are very natural in a regression context. Firstly, $f^0(P)(Y \leq t) = P(Y \leq t)$ for all $P$ and all $t$, so there is an *identity* transformation $f^0$, permitting us to represent the idea that the response distribution does not vary with this covariate (or a group of covariates). Secondly, we insist that for all real $v, v'$ and all $P$ and $t$, we have $[f^{v'}\{f^v(P)\}](Y \leq t) = P(Y \leq t - v - v') = P\{Y \leq t - (v + v')\} = f^{v+v'}(P)(Y \leq t)$. This second flow property captures the idea of homogeneous *accumulation* of modifications to distributions, for instance to express a dose-response relationship.

The second flow property also motivates the convenient superscript notation for the subordinate argument of $f$—the flow index $v$—mirroring for function composition the addition of exponents in ordinary multiplication. Though its meaning is intuitively clear, this property of flows can be expressed even more succinctly if we write function application unbracketed and on the right, so that for example $Pf^v$ means precisely $f^v(P)$. The second flow property then insists simply that $Pf^v f^{v'} = Pf^{v+v'}$ for all $v, v'$, and $P$.

This location shift flow is indexed by a real number $v$. Depending on the range of the response variable $Y$ and also on the specific flow in question, it may be naturally indexed by a scalar, a tuple, or a function. We insist only that the index $v$ belong to a vector space, which for the time being we denote generically by $\mathbb{V}$.

Definition 3.1 (Flow). *Let $\mathcal{M}$ be the set of signed finite measures on the measurable space $(\Upsilon, \mathcal{Y})$, and let $\mathbb{V}$ be a vector space with identity element 0. Then $f : \mathcal{M} \times \mathbb{V} \to \mathcal{M}$ is a* flow *if it satisfies the following properties:*

*(a)* $f^0 = f(\cdot, 0)$ *is the identity transformation.*
*(b)* $f^v f^{v'} = f(f(\cdot, v), v') = f(\cdot, v + v') = f^{v+v'}$ *for all* $v, v' \in \mathbb{V}$.

Another term for a flow is a *dynamical system* (Brown, 2018), wherein the effect size parameter $v$ typically represents the passage of time. Flows characterize a family of transformations that describes a natural, continuous equivalent of function iteration. In words, the first flow property says that if $f$ is 'iterated' zero times, it is the identity function, and an input law $P$ is unchanged by application of $f$. The second property guarantees that the result of 'applying the function' twice (say, so $v = 2$) and then 'applying it' three times (say, so $v' = 3$) is the same as 'applying the function' five times. Implicitly, it also provides a definition for what it means to 'apply the function' $f$ one and a half times, $e$ times, or minus three times. Were $\mathbb{V}$ a function space, the flow could tell us what it means to 'apply the function' $f$ a vector $v : t \mapsto t^2$ (say) 'number of times', so to speak.

The flow properties equip the set $f^{\mathbb{V}} = \{f^v : v \in \mathbb{V}\}$ with a non-trivial algebraic structure, and one that is not typically present in arbitrary families of transformations. The family of functions making up a flow inherits from $\mathbb{V}$ the properties of a group, with function composition as the group operation. The iterative argument $v$ is an index that respects the group structure of the vector space

$\mathbb{V}$: each 'iteration' $v$ of the function has a unique inverse $-v$, satisfying $f^v f^{-v} = f^0 =$ identity. So-equipped with composition as its group operation, $f^{\mathbb{V}}$ is a subgroup of the group of *all* transformations of $\mathcal{M}$. Although function composition is not in general commutative, flows are in fact abelian groups, a property again inherited from the vector space $\mathbb{V}$.

Looked at another way, a flow $f$ is a group homomorphism, identifying each element $v \in \mathbb{V}$ with a corresponding transformation $f^v$ in $f^{\mathbb{V}}$. If this identification is bijective, the group homomorphism is a vector space isomorphism: the elements of $f^{\mathbb{V}}$ are transformations of $\mathcal{M}$, and the 'addition' operation in $f^{\mathbb{V}}$ is function composition. The operations of 'addition' (function composition) and 'scalar multiplication' (function iteration) in $f^{\mathbb{V}}$ exactly mirror vector addition and scalar multiplication in $\mathbb{V}$, in whatever way these operations in $\mathbb{V}$ are defined. Whenever the group homomorphism of $\mathbb{V}$ and $f^{\mathbb{V}}$ defined by $f$ is bijective—as indeed it is in most examples considered here—$\mathbb{V}$ and $f^{\mathbb{V}}$ are algebraically indistinguishable. However, even though the vector space $f^{\mathbb{V}}$ of transformations is our primary interest, we nevertheless retain $v \in \mathbb{V}$ as a convenient index with which to refer to specific transformations $f^v \in f^{\mathbb{V}}$.

In permitting the flow index space $\mathbb{V}$ to be an arbitrary vector space, Definition 3.1 is slightly more general than the one typically encountered, wherein flows are indexed simply by the real line. This abstraction is essential: flows indexed by functions allow us to specify the Cox flow and Aalen flow, for example, which we visited in Section 1 and return to in Section 4.2.3. Having taken this small liberty, we now take another, and mention that it can sometimes be of interest to index flows by spaces more general still, such as Lie groups (see, for example, Lee, 2013, pp. 150 sqq.)—which may not even be abelian. One such instance arises naturally when studying the accelerated failure time model for survival data, which we discuss in Section 4.1.1. The idea of composing transformations can also readily be generalized to specific non-flow families, such as the switch relative risk (van der Laan et al., 2007). We return briefly to this point in Section 6.

For the most part, however, we shall content ourselves to study flows indexed by vector spaces. Restricting attention to such flows simplifies our exposition, and leads to model properties that would not hold for compositions of less restricted classes of transformations.

### 3.2.2. *Generating functions*

Regressions by composition are parametric models, but their component flows are semiparametric in the sense that they must be defined for all measures $P \in \mathcal{M}$, and not (say) only on the set of laws corresponding to normal distributions. This is needed because algebraic properties of transformations (such as linearity) are determined by their behaviour on the whole of $\mathcal{M}$, not just on a small parametric subset. One implication of the admittedly strong requirement to specify transformations on all of $\mathcal{M}$ is that it is not always straightforward to describe how various existing regression models should be extended to accommodate it. As a simple example, many regression models allow covariates to 'change the mean' of the distribution of $Y$ while remaining within the same parametric family. How should such a notion be extended to accommodate arbitrary input measures $P$ that do not necessarily belong to this parametric family?

There may even be more than one possible extension (see Section 4 for an example). However, in some cases, different ways of characterizing the laws in $\mathcal{M}$ can make particular flow generalizations of existing regression models seem especially natural. For instance, if we represent a law $P$ by its cumulative distribution function $\hat{P}\colon t \mapsto P(Y \leq t)$, then the location shift flow can be defined by $(\hat{P} f^v)(t) = \hat{P}(t - v)$ and does indeed 'change the mean', increasing it (and every quantile) by $v$.

However, this operation is well-defined for all measures $P \in \mathcal{M}$, including those without a mean!

By allowing $f$ to act on the cumulative distribution function $\hat{P}$, and to output a cumulative distribution function $\hat{P}f$, we have deliberately blurred the distinction between the measure $P$ and its distribution function $\hat{P}$. This may be justified because, of course, the two objects are in an important sense equivalent: the latter may be identified with the former, but 'wearing a funny hat'. Rather more formally, the map $\phi\colon P \mapsto \hat{P}$ defines an isomorphism $\mathcal{M} \cong \phi(\mathcal{M}) = \hat{\mathcal{M}}$ of vector spaces. In our distribution function example, the isomorphism $\phi$ identifies $P$ with its cumulative distribution function $\phi(P)\colon \mathbb{R} \to [0,1]$ defined by $\phi(P)(t) = P(Y \leq t)$. The probability-respecting vector space structure of $\mathcal{M}$ means that $\phi$ is indeed an invertible linear map, because $\phi(aP + bQ)(t) = (aP + bQ)(Y \leq t) = aP(Y \leq t) + bQ(Y \leq t) = a\phi(P)(t) + b\phi(Q)(t)$ for measures $P, Q$ and real numbers $a, b$ and $t$.

We will often write $\hat{P}$ for $\phi(P)$; both notational choices are intended to evoke a very widely-applicable isomorphism, namely the map $\phi$ that takes $P$ to its characteristic function $\phi(P)\colon t \mapsto P\{\exp(itY)\} = \hat{P}(t)$, where $i = \sqrt{-1}$. Nevertheless, we will employ the same notation $\hat{P}$ for various other generating functions, including the survivor function, the probability generating function and the moment generating function; we endeavor to make it clear from the context which particular characterization $\hat{P}$ of $P$ we have in mind. For laws on finite sets $\Upsilon = \{\upsilon_1, , \ldots, \upsilon_N\}$ it can sometimes be helpful to identify $P$ with the point $\phi(P) = (P(Y = \upsilon_1), \ldots, P(Y = \upsilon_N))$ in Euclidean space $\mathbb{R}^N$: this too characterizes an isomorphism of vector spaces. If $Y$ is binary, another convenient choice is $\phi(P) = (P(\Omega), P(Y = 1))$, because holding $P(\Omega) = 1$ provides a direct connection to the familiar l'Abbé plot; see Section 4.2.1 for an example of this representation.

Later in the paper we will give further instances of flows on different spaces $\mathcal{M}$. However, to exhibit a case with $\mathbb{V} \neq \mathbb{R}$, recall the family of transformations corresponding to the Cox model (and introduced informally in Section 2): we now show that this family is indeed a flow. Let $\mathbb{V}$ be the set of positive-valued functions of time $[0, \infty) \to (0, \infty)$. Equip $\mathbb{V}$ with the group operation of pointwise multiplication (denoted by juxtaposition as $vv'$), so that the function $t \mapsto 1$ is its identity element. The Cox flow $f$ transforms a survivor function $\hat{P}$ as

$$(\hat{P}f^v)(t) = \prod_{s=0}^{t} \{\mathrm{d}\hat{P}(s)\}^{v(s)}$$

where again $\mathrm{d}\hat{P}(s)$ is a multiplicative increment. Then $(Pf^{t\mapsto 1})(t) = \prod_{s=0}^{t}\{\mathrm{d}\hat{P}(s)\}^1 = \hat{P}(t)$, and, for $v, v' \in \mathbb{V}$, $(Pf^v f^{v'})(t) = \prod_{s=0}^{t}[\{\mathrm{d}\hat{P}(s)\}^{v(s)}]^{v'(s)} = \prod_{s=0}^{t}\{\mathrm{d}\hat{P}(s)\}^{v(s)v'(s)} = (\hat{P}f^{vv'})(t)$ as required. This example illustrates that the group operation in $\mathbb{V}$ need not be addition (or even addition of functions): the Cox model is naturally multiplicative.

### 3.2.3. *Linear predictors*

Although the flow index space $\mathbb{V}$ may in principle be infinite-dimensional—a function space, for example, as in the case of the Cox flow described at the end of the previous section—we shall in practice work with finite-dimensional vector spaces. For some $q \geq 1$, we assume $\mathbb{V}$ has a user-specified basis $(v^1, \ldots, v^q)$, each $v^k$ being an element of $\mathbb{V}$. If $\mathbb{V}$ is a function space, splines offer one possible way to construct such a basis (Schoenberg, 1946a, 1946b).

How does covariate information map into the vector space $\mathbb{V}$ to distinguish a specific transformation in $f^{\mathbb{V}}$? To achieve this in a flexible way, we define an $\mathcal{F}$-measurable random variable $X\colon \Omega \to \mathbb{U}$ that embeds covariate comparisons (with either an explicit or a notional reference

value) into a $p$-dimensional vector space $\mathbb{U}$, and relate it to the $q$-dimensional $\mathbb{V}$ via a $(p \times q)$-dimensional linear map $\theta \in \Theta$. Maps of the form $\theta\colon \mathbb{U} \to \mathbb{V}$ are the only components of a regression by composition that are estimated from data, and form the crucial links between covariate contrasts (as quantified within the embedding space $\mathbb{U}$) and transformations of the outcome distribution (as quantified within the flow index space $\mathbb{V}$). In particular, the $(p \times q)$-dimensional vector space $\Theta$ contains the zero map, which expresses the idea of 'no dependence on this covariate'; it maps any real $p$-vector of covariate comparisons to the identity element in $\mathbb{V}$, and thence to the identity element $f^0$ of the flow $f$. If $\mathbb{U} = \mathbb{R}^p$ for some $p$, as will often be the case, then embedding covariate contrasts in Euclidean space is just a more formal way of specifying a standard model matrix (in linear regression, say), and indeed software tools used to construct so-called *design matrices* are very convenient for creating suitable $p$-vector embeddings of covariates.

For the purposes of estimation, we associate a map $\theta\colon \mathbb{U} \to \mathbb{V}$ with its matrix representation $[\theta]$ with respect to bases $(u^1, \ldots, u^p)$ of $\mathbb{U}$ and $(v^1, \ldots, v^q)$ of $\mathbb{V}$. The real-valued components $[\theta]^k_l$ of the $p \times q$ matrix $[\theta]$ are defined in terms of the images $u^k\theta$ of the basis vectors $u^k$ under $\theta$. Employing Einstein's summation convention, every such image satisfies an equation of the form $u^k\theta = [\theta]^k_l v^l$, where $[\theta]^k_l$ is the element in row $k$ and column $l$ of the matrix $[\theta]$, and summation over $l$ is implicit. We emphasize again that both scalar multiplication (of $[\theta]^k_l \in \mathbb{R}$ by $v^l \in \mathbb{V}$) and summation (over $l$) take place within the vector space $\mathbb{V}$, and need not be real multiplication and addition. Given real coefficients $x_k$ defining a generic element $x = x_k u^k$ of $\mathbb{U}$ (here with addition and scalar multiplication as defined in $\mathbb{U}$), we may write $x\theta = (x_k u^k)\theta$, which by linearity of $\theta$ is equivalently $x_k(u^k\theta) = x_k([\theta]^k_l v^l) = (x_k[\theta]^k_l)v^l$.

We operationalize the estimation of the linear map $\theta$ via two 'convenience' functions. For likelihood optimization, we shall want to express derivatives compactly and functionally, for example as $\ell^{(\Theta)}$ rather than $\mathrm{d}\ell/\mathrm{d}\theta$. To this end, we define the function $\beta\colon \mathbb{U} \times \Theta \to \mathbb{V}$ to satisfy $\beta(x, \theta) = x\theta$, the right-hand-side being the application of the *function* $\theta$ to the vector $x$. We then define the *linear predictor* $\eta\colon \Omega \times \Theta \to \mathbb{V}$ to be the composition of a $\mathbb{U}$-valued random variable $X$ with $\beta$; that is, $\eta = X\beta$, where necessarily the output of $X$ is attached to the $\mathbb{U}$-valued input of $\beta$, the $\Theta$-valued argument of $\beta$ being left dangling (so to speak) along with the $\Omega$-valued argument of $X$.

Although we construct the linear predictor $\eta$ as a function, rather than a numeric variable, it is nevertheless analogous to the linear predictors used in generalized linear models, and indeed completely equivalent when their corresponding flows are employed. A subtle but significant difference is that, in a regression by composition, the linear predictor is not necessarily linked to a transformation of the mean of $Y$, but to any 'size' index of a flow. This link is formalized by a further function composition: we define the random, parametrized transformation $F\colon \mathcal{M} \times \Omega \times \Theta \to \mathcal{M}$ as $F = f^\eta = \eta f$, where again necessarily the output of $\eta$ is connected to the $\mathbb{V}$-valued 'effect size' argument of the deterministic flow $f$. Regression by composition stacks together several of these estimable, covariate-dependent transformations $F$ in a manner that we can now (finally!) describe.

### 3.2.4. *Sequential transformation of laws*

In modelling the conditional law $\mathbb{P}$, we shall require a notional ordering of the explanatory information $\mathcal{F}$. In many familiar settings, some or all model components are in fact invariant to the choice of ordering, so that no special knowledge of temporal or logical ordering is required. Nevertheless, there exist models within our framework for which order does matter, and in such

cases this will ideally arise from subject-matter considerations, and in particular from causal reasoning about the potential impacts of covariates on the response $Y$: the impacts of genetics logically precede epigenetic effects, and developmental or childhood exposures in turn precede effects of interventions experienced in adulthood.

Covariate ordering is captured by a finite filtration $(\mathcal{F}_j) = (\mathcal{F}_1, \ldots, \mathcal{F}_m)$ of $\mathcal{F}$, with $\mathcal{F}_m = \mathcal{F}$ for some positive integer $m$. The filtration $(\mathcal{F}_j)$ represents increasing covariate information (see, for example, Andersen et al., 1996, pp. 59 sqq.): in the Cox–Aalen example of Section 2, we set $\mathcal{F}_1 = \sigma(\emptyset)$, $\mathcal{F}_2 = \sigma(A)$ and $\mathcal{F}_3 = \sigma(A, G)$. To each pair $(j-1, j)$ we associate a flow $f_j$ that describes how the corresponding increment in covariate information is to be incorporated into the model. A novelty of regression by composition is that the filtration $(\mathcal{F}_j)$ need not be *strictly* increasing: this allows, for example, the same covariate information (treatment, say) to change the distribution of $Y$ in two qualitatively different ways, perhaps first altering its conditional mean and then its conditional variance. Equally, more than one covariate can enter at a given stage, and indeed in most existing regression models $m = 1$ and $\mathcal{F}_1$ is generated by *all* covariates, parametrized through a single linear predictor $\eta_1$ and a single flow $f_1$.

We define a corresponding sequence $(X_j)$ of embedded covariate contrasts, where each $X_j$ is $\mathcal{F}_j$-measurable and takes values in a $p_j$-dimensional vector space $\mathbb{U}_j$. These $X_j$ might depend only on the 'new' information in $\mathcal{F}_j$, but could also incorporate so-called interaction terms with covariates available in $\mathcal{F}_{j-1}$. The embedded covariate comparisons $X_j$ are then mapped to a $q_j$-dimensional vector space $\mathbb{V}_j$ via the local linear predictor $\eta_j = X_j\beta_j$, which when composed with its associated flow $f_j$ yields the transformation $F_j = f_j^{\eta_j} = \eta_j f_j$ of $\mathcal{M}$.

The conditional law of $Y$ changes as we progressively incorporate the covariate information $(\mathcal{F}_j)$. This evolution is captured through a sequence of random laws $(\mathbb{P}_j)$ satisfying the recurrence relation

$$\mathbb{P}_j = \mathbb{P}_{j-1}F_j$$

for $j = 1, \ldots, m$. In words, $\mathbb{P}_j$ is the composition of $\mathbb{P}_{j-1}$ (an $\mathcal{F}_{j-1}$-measurable parametrized law) and the (random, parametrized) transformation $F_j$. Although the flows $f_j$ are user-specified and deterministic, the transformations $F_j$ are random because they depend on covariates and so, like $(X_j)$, the sequence $(F_j)$ is also a stochastic process adapted to $(\mathcal{F}_j)$. Given some initial, non-random law $\mathbb{P}_0$ on $(\Upsilon, \mathcal{Y})$, we may express $\mathbb{P}_m$, our parametric model for $\mathbb{P}$, as

$$\mathbb{P}_m = \mathbb{P}_0 F_1 \cdots F_m,$$

where here again unbracketed functions on the right denote function composition, so that $F_1$ is applied first, then $F_2$, and so on. The model $\mathbb{P}_m$ is an $\mathcal{F}$-measurable, parametrized law $\mathcal{Y} \times \Omega \times \Theta_1 \times \cdots \times \Theta_m \to [0, 1]$, and has the ambition that, for some $(\theta_1, \ldots, \theta_m)$ belonging to $\Theta_1 \times \cdots \times \Theta_m$,

$$\mathbb{P} = \mathbb{P}_m(\theta_1, \ldots, \theta_m).$$

For fixed $\theta_1, \ldots, \theta_m$, the sequence $(\mathbb{P}_j)$ is a measure-valued stochastic process adapted to $(\mathcal{F}_j)$. The model $\mathbb{P}_m$ is constructed as a finite composition of transformations $F_j$, its component flows $f_j$ themselves being essentially *continuous* compositions; hence the term *regression by composition*.

### 3.3. Properties of transformations

We claim that the properties of flows $f$, or of their component transformations $f^v$, make clear certain local features of a regression by composition that may be important either in theory or

in particular applications. When we say that a *flow* $f$ has a particular property, we will usually mean by this that the property should apply to its constituent transformations $f^v$ for all indices $v \in \mathbb{V}$. Many properties we describe in this section can also hold (or not) for compositions $f^v g^w$ of transformations or combinations $fg$ of flows.

### 3.3.1. *Closure*

We begin with the simpler properties of transformations, whose definitions and implications for modelling are more straightforward. In Section 2 we alluded to the fact that the additive (Aalen, 1989) flow can sometimes map valid input survivor functions to output functions that cannot be survivor functions (for instance, to a function that is not monotonic decreasing). This corresponds directly to mapping valid probability laws to measures that are not probability laws (for instance, those placing negative mass on some events, or having non-unit total mass). We formalize this important property using the standard definition of a set being closed under a specific operation.

Definition 3.2 (Closure). *A transformation $f^v$ is* closed *on a set $M \subset \mathcal{M}$ of measures if $Pf^v \in M$ for all $P \in M$, or more compactly if $Mf^v \subseteq M$. For brevity we shall sometimes simply say that the transformation $f^v$ is* closed*, with the set $M$ unspecified, by which we mean that $f^v$ is closed on $M = \mathcal{P}$, the set of all probability laws on $(\Upsilon, \mathcal{Y})$. A flow $f$ is closed if and only if* all *its component transformations $f^v$ are closed.*

The predominance of logistic regression over linear probability models, or of the Cox model over the Aalen model, can in part be explained by an understandable preference among the statistical community for closed flows. For a counterargument with particular reference to binary regression, see for example Hellevik (2009). In any event, the importance of closure as a property of a flow can be application-specific.

### 3.3.2. *Constraints*

Another algebraic feature of transformations concerns whether it is possible for a law to get 'stuck': in other words, if applying a non-identity transformation $f^v$ leaves certain laws unchanged. This relates to the *zero constraints* of Deeks (2002), who uses the term to mean that at certain points $P \in \mathcal{M}$ a particular intervention is constrained by the model to have *zero benefit or harm* across a whole family of transformations. A good example is the odds ratio model (that is, logistic regression), which is "constrained to predict absolute benefits of zero both when the control group event rate is 0 per cent and when it is 100 percent" (Deeks, 2002, p. 1584).

Definition 3.3 (Fixed point). *The law $P$ is a* fixed point *of the transformation $f^v$ if $Pf^v = P$. The law $P$ is a fixed point of a flow $f$ if it is a fixed point for all its constituent transformations $f^v$.*

In practice, fixed points of transformations or flows are unlikely actually to be reached by a model-fitting algorithm, but can nevertheless indicate a certain rigidity *in the vicinity of* the fixed point that may have important consequences. For instance, it is difficult for logistic regression (with two fixed points) to capture the effect of an intervention that is of substantial benefit for people at high risk of experiencing the event of interest, but is of diminishing value as the untreated risk decreases towards zero. By contrast, a log-linear or relative risk model (having only one fixed point) describes such an effect in a straightforward manner.

Fixed points can also have implications for the assessment of model uncertainty. Confidence intervals (or credible intervals, or likelihood intervals) around predictions located well away from the bulk of the data can be apparently precise or vague depending on whether there is, or is not, assumed to be a fixed point in the vicinity of the prediction being made.

### 3.3.3. *Exhaustion*

Exhaustion is exclusively a property of a flow, not of an individual transformation. It characterizes a flow as being locally 'nonparametric', in the sense that *any* distribution can be reached from any starting point.

DEFINITION 3.4 (EXHAUSTION). *A flow $f$ is* exhaustive *if, for almost all $P, Q \in \mathcal{P}$, there exists $v \in \mathbb{V}$ such that $Q = Pf^v$.*

The *almost all* caveat makes allowance for a countable number of fixed points. Any such fixed points can usually be sidestepped by choosing an initial law $\mathbb{P}_0$ that is *not* a fixed point. Semiparametric models, such as those arising from Cox (1972) regression, can be interpreted as employing exhaustive flows as the first 'layer' in a regression by composition construction.

Exhaustion should not be confused with saturation, which is a standard property of a covariate contrast embedding $X$ in any modelling that employs a linear predictor. An embedding $X$ is *saturated* if the dimension $p$ of the embedding equals the number of distinct possible values that can be assumed by the corresponding covariate(s). A standard example would be a model layer that permits each of three treatment groups to have a different mean. Exhaustion and saturation are therefore both indicative of model flexibility, but in rather different senses: one characterizes a flexible outcome distribution, and the other flexible conditioning.

### 3.3.4. *Equivalence*

Two flows are equivalent if they represent the same family of transformations; in other words, equivalent flows span the same (local) model space. To make this statement more precise, recall that $f^{\mathbb{V}}$ means the set $\{f^v : v \in \mathbb{V}\}$ of transformations of $\mathcal{M}$ equipped with the operation of function composition, and similarly $g^{\mathbb{W}} = \{g^w : w \in \mathbb{W}\}$ for some vector space $\mathbb{W}$.

DEFINITION 3.5 (EQUIVALENCE). *Two flows $f$ and $g$ with corresponding index spaces $\mathbb{V}$ and $\mathbb{W}$ are* equivalent *if $f^{\mathbb{V}} = g^{\mathbb{W}}$.*

A simple example suffices to illustrate this point. Logistic regression can be parametrized in terms of a (multiplicative) odds ratio, taking values in $(\mathbb{R}_+, \times)$ or, more usually, as an (additive) log odds ratio, taking values in $(\mathbb{R}, +)$. The two are fundamentally the same, and only matters of computational convenience or interpretability lead us to choose one over the other.

### 3.3.5. *Commutativity*

Two transformations $f^v$ and $g^w$ commute if $f^v g^w = g^w f^v$. Commutativity of flows is a slightly more subtle concept, and has at least two natural meanings, one strictly stronger than the other. Roughly, weak commutativity of two flows means that the same global model arises irrespective of the order in which they are deployed. Strong commutativity means that any pairwise

combination of the flows' component transformations commute, and ensures that the composition of the two flows is itself a flow. In what follows, we denote by $f^{\mathbb{V}}g^{\mathbb{W}}$ the set of transformations $\{f^v g^w : u \in \mathbb{V}, v \in \mathbb{W}\}$.

Definition 3.6 (Commutativity). *The flows $f$ and $g$* commute weakly *if $f^{\mathbb{V}}g^{\mathbb{W}} = g^{\mathbb{W}}f^{\mathbb{V}}$. The flows $f$ and $g$* commute strongly *if $f^v g^w = g^w f^v$ for all $v \in \mathbb{V}, w \in \mathbb{W}$.*

Strong commutativity implies weak commutativity, and a standard theorem of group theory tells us that weak commutativity is equivalent to the statement that $f^{\mathbb{V}}g^{\mathbb{W}}$ is a subgroup (though not necessarily abelian) of the whole group of transformations of $\mathcal{M}$. More importantly for our purposes, weak commutativity implies that the model space spanned by the composition of the two flows is independent of their ordering, and that the two orderings are equivalent in the sense of Definition 3.5. Strong commutativity gives us more:

Theorem 3.1. *Let $f$ and $g$ be strongly commuting flows, with index spaces $\mathbb{V}$ and $\mathbb{W}$, respectively. Then the composition $fg$ is a flow, with index space $\mathbb{V} \times \mathbb{W}$.*

Proof. Certainly $f^0 g^0$ (the identity transformation) is in $f^{\mathbb{V}}g^{\mathbb{W}}$. To confirm that the second, aggregating flow property holds, consider two generic elements $f^v g^w$ and $f^{v'}g^{w'}$ of $f^{\mathbb{V}}g^{\mathbb{W}}$. We have that $f^v g^w f^{v'} g^{w'} = f^v f^{v'} g^w g^{w'} = f^{v+v'} g^{w+w'}$, so $fg$ is a flow under componentwise vector addition in $\mathbb{V} \times \mathbb{W}$. □

A consequence of these facts is that a full regression by composition $f_1 \cdots f_m$ need not be a flow, or indeed even a subgroup of the group of transformations of $\mathcal{M}$. The model will be (locally) invariant to order if flows commute weakly; strongly commuting flows can often helpfully be thought of as a single flow. One mathematically trivial but nevertheless important quality of any flow is that it strongly commutes with itself. As a result, two covariates acting through the same flow $f$ but in adjacent model layers can equivalently be aggregated into a single layer, if desired.

### 3.3.6. *Collapsibility*

To motivate our general definition of collapsibility, return again to the Cox–Aalen example of Section 2. Imagine that the conditional distributions of survival time $Y$ given participant age $A$ (via an Aalen flow) and treatment group $G$ (via a Cox flow) have now been estimated from a large trial in which treatment was randomly assigned. The transformation associated with the Cox flow index $\eta_3$ is an estimated *conditional* treatment effect (given age) and indeed may even depend on age through what might traditionally be called interaction terms. However, perhaps a *marginal* treatment effect is also of interest. What, if anything, may we deduce from our fitted Cox-Aalen model about a marginal comparison of treated and untreated groups?

Let $\mathbb{P}_2$ denote as before the estimated conditional control group law of $Y$ given age $A$, and $\mathbb{P}_3$ the final fitted regression by composition: that is, the estimated conditional law of $Y$ given age $A$ and treatment group $G$. Writing $\mathcal{G} = \sigma(G)$, we are interested in computing $\mathbb{E}(\mathbb{P}_3 \mid \mathcal{G}) = \mathbb{E}\{f_3(\mathbb{P}_2, \eta_3) \mid \mathcal{G}\}$. Can we somehow relate this target quantity to the readily available $\mathbb{E}(\mathbb{P}_2 \mid \mathcal{G}) = \mathbb{E}(\mathbb{P}_2)$ (a non-random marginal law, since treatment is independent of age as a result of randomization), our deterministic Cox flow $f_3$, and any available bounds on the distribution of the age-related treatment effect $\eta_3$ in the treatment group? (In the control group, $\eta_3 = 0$ by construction.) We call flows *collapsible* where use of boundaries is possible in this way, and make the following definition:

Definition 3.7 (Collapsibility). *Let $P$ be an $\mathcal{F}$-measurable random law on $(\Upsilon, \mathcal{Y})$ and $\eta$ an $\mathcal{F}$-measurable random element of the vector space $\mathbb{V}$. Let $\mathcal{G} \subseteq \mathcal{F}$, and suppose that $\eta$ belongs to a $\mathcal{G}$-measurable random subset $H \subseteq \mathbb{V}$ almost surely. We say that the flow $f$ is collapsible over $H$ if*

$$\mathbb{E}\{f(P, \eta) \mid \mathcal{G}\} = f\{\mathbb{E}(P \mid \mathcal{G}), h\} \text{ for some } h \in \operatorname{Conv} H$$

*for all measures $\mathbb{E}$ and all such $P, \eta$, and $\mathcal{G}$, where* Conv *denotes the convex hull of a set.*

In the Cox-Aalen example, if the Cox flow $f$ changes an age-specific control law $P$ into a corresponding age-specific treated law $f(P, \eta)$, it does so by raising to some estimated age-specific positive exponent $\eta$ the proportion of individuals surviving each small increment of time. For the Cox flow to be collapsible on a set of such age-specific exponents, say $H = [2, 5]$, we would require that an exponent $2 \leq h \leq 5$ exists such that the transformation between the all-age control law $\mathbb{E}(P)$ and the all-age treated law $\mathbb{E}\{f(P, \eta)\}$ is again of the Cox form $f$, with exponent $h$.

Our collapsibility condition is similar to existing definitions (e.g. Greenland et al., 1999; Huitfeldt et al., 2019) but differs in one important respect. Instead of relating *functional* contrasts between conditional and marginal distributions, we compare distributions directly within the space $\mathcal{M}$. As in Huitfeldt et al. (2019), the crucial point is still whether the marginal distributions $\mathbb{E}\{f(P, \eta) \mid \mathcal{G}\}$ are bounded by transformations $f^h$ of $\mathbb{E}(P \mid \mathcal{G})$ in the convex hulls $\operatorname{Conv} f^H$. An advantage of the present formulation is that it obviates the "technical problem" highlighted by Greenland et al. (1999, 38[38) wherein marginal and conditional models may not have the same functional form.

If a flow $f$ is collapsible, and if in fact $\eta$ is $\mathcal{G}$-measurable (in our Cox-Aalen example, if the conditional treatment effect size $\eta_3$ does not, after all, depend on age $A$), then $H$ is a $\mathcal{G}$-measurable random singleton set, and the convex hull in the definition collapses to a single point:

$$\mathbb{E}\{f(P, \eta) \mid \mathcal{G}\} = f\{\mathbb{E}(P \mid \mathcal{G}), \eta\}.$$

Greenland et al. (1999) call this latter condition *strict collapsibility*.

Before proving Theorem 3.2, which provides conditions under which a flow $f$ is collapsible, we require three standard definitions. Although the standard definitions are much broader in scope, we articulate them here as they pertain to transformations and measures.

Definition 3.8 (Linearity). *A transformation $f^v$ is* linear *if and only if*

$$f^v(aP + bQ) = a f^v(P) + b f^v(Q)$$

*for all scalars $a, b$ and all measures $P, Q$. A flow $f$ is linear if all its consituent transformations $f^v$ are linear.*

Definition 3.9 (Affinity). *A transformation $f^v$ is* affine *if and only if it can be written in the form $f^v = g + c$, where function addition is componentwise, $g$ is linear and $c$ is a constant function. A flow $f$ is affine if all its constituent transformations $f^v$ are affine.*

Definition 3.10 (Convexity). *A region $R \in \mathcal{M}^2$ is* convex *if, for any pair of points $(P, Q)$ and $(P', Q')$ in $\mathcal{M}^2$, the line segment joining $(P, Q)$ to $(P', Q')$ lies entirely within $R$.*

THEOREM 3.2. *A flow is collapsible over a $\mathcal{G}$-measurable set $H \subseteq \mathbb{V}$ if and only if the set $\{(P, Pf^h) : P \in \mathcal{P}, h \in \operatorname{Conv} H\}$ is convex.*

COROLLARY 3.2.1. *A flow is strictly collapsible if and only if it is affine.*

Proof. In this proof, we let $\mathbb{P}$ denote an $\mathcal{F}$-measurable random law (taking values in $\mathcal{P}$), while $P$ is a generic deterministic law. For the implication in one direction, suppose that $\{(P, Pf^h) : P \in \mathcal{P}, h \in \operatorname{Conv} H\}$ is convex. Consider $\mathbb{E}\{(\mathbb{P}, \mathbb{P}f^\eta) \mid \mathcal{G}) = \{\mathbb{E}(\mathbb{P} \mid \mathcal{G}), \mathbb{E}(\mathbb{P}f^\eta \mid \mathcal{G})\}$. By convexity of expectation,

$$\begin{aligned}\mathbb{E}\{(\mathbb{P}, \mathbb{P}f^\eta) \mid \mathcal{G}\} &\in \operatorname{Conv}\{(P, Pf^h) : P \in \mathcal{P}, h \in H\} \\ &\subseteq \operatorname{Conv}\{(P, Pf^h) : P \in \mathcal{P}, h \in \operatorname{Conv} H\} \\ &= \{(P, Pf^h) : P \in \mathcal{P}, h \in \operatorname{Conv} H\}\end{aligned}$$

by the assumed convexity of the latter set. Therefore $\{\mathbb{E}(\mathbb{P} \mid \mathcal{G}), \mathbb{E}(\mathbb{P}f^\eta \mid \mathcal{G})\} = (P, Pf^h)$ for some $P \in \mathcal{P}$ and some $h \in \operatorname{Conv} H$. But this $P$ must be precisely $\mathbb{E}(\mathbb{P} \mid \mathcal{G})$, so $\mathbb{E}(\mathbb{P}f^\eta \mid \mathcal{G}) = \mathbb{E}(\mathbb{P} \mid \mathcal{G})f^h$ for some $h \in \operatorname{Conv} H$, as required for collapsibility of $f$ over $H$.

For the converse, consider a flow $f$ and a $\mathcal{G}$-measurable set $H$ such that $\{(P, Pf^h) : P \in \mathcal{P}, h \in \operatorname{Conv} H\}$ is *not* convex. Then there exist $\mathcal{G}$-measurable points $(P_\rightarrow, Q_\rightarrow)$, $(P_*, Q_*)$, and $(P_\leftarrow, Q_\leftarrow)$ such that $(P_\rightarrow, Q_\rightarrow), (P_\leftarrow, Q_\leftarrow) \in \{(P, Pf^h) : P \in \mathcal{P}, h \in \operatorname{Conv} H\}$, $(P_*, Q_*) \in \operatorname{Conv}\{(P_\rightarrow, Q_\rightarrow), (P_\leftarrow, Q_\leftarrow)\}$ but $(P_*, Q_*) \notin \{(P, Pf^h) : P \in \mathcal{P}, h \in \operatorname{Conv} H\}$. Now let the measure $\mathbb{E}$ and the random variables $\mathbb{P}$ and $\eta$ be so defined that the pair $(\mathbb{P}, \mathbb{P}f^\eta)$ of random measures takes values in the two-point set $\{(P_\rightarrow, Q_\rightarrow), (P_\leftarrow, Q_\leftarrow)\}$ and has expectation $\mathbb{E}\{(\mathbb{P}, \mathbb{P}f^\eta) \mid \mathcal{G}) = (P_*, Q_*)$; this is always possible because $(P_*, Q_*) \in \operatorname{Conv}\{(P_\rightarrow, Q_\rightarrow), (P_\leftarrow, Q_\leftarrow)\}$. Then $\mathbb{E}(\mathbb{P}f^\eta \mid \mathcal{G}) = Q_*$, and $Q_*$ does not equal $\mathbb{E}(\mathbb{P} \mid \mathcal{G})f^h$ for any $h \in \operatorname{Conv} H$. We conclude that the flow $f$ is not collapsible over $H$.

For the corollary, assume that $\eta$ is $\mathcal{G}$-measurable, so that $H = \{\eta\}$ is a singleton set. Then $\{(P, Pf^h) : P \in \mathcal{P}, h \in \operatorname{Conv} H\} = \{(P, Pf^\eta) : P \in \mathcal{P}\}$, describing a segment of a curve in the space $\mathcal{M}^2$. Such a curve segment is a convex set if and only if $f^\eta$ is an affine function on $\mathcal{P}$. □

As we have already noted, the Cox model is not collapsible, so our hoped-for bounding of marginal effects described at the start of this section is not possible in this case. However, in Section 4 we do provide several examples of flows that are collapsible.

Only affine flows can be strictly collapsible, but beyond strict collapsibility the situation becomes more complicated. Some affine flows are collapsible over all subsets $H$ of their index space $\mathbb{V}$, while others are not collapsible except over singleton sets (strict collapsibility). Conversely, non-affine flows cannot be collapsible over all subsets of $\mathbb{V}$—because they do not collapse over singleton sets—but some can nevertheless be collapsible over certain larger subsets $H \subseteq \mathbb{V}$.

These points can most easily be understood using a somewhat abstract version of the l'Abbé plot (L'Abbé et al., 1987). Instead of plotting the probabilities $(Pf^v)(Y = 1)$ against $P(Y = 1)$ in the space $[0, 1]^2$, we instead imagine plotting the laws $Pf^v$ against $P$ in the space $\mathcal{M}^2$. Of course, $\mathcal{M}$ has a minimum of two dimensions itself, meaning that this plotting must be somewhat notional, although still helpful conceptually.

A one-dimensional affine flow with a fixed point at $P$ in the interior of $\mathcal{P}$ must in some sense 'hinge' at this point. It is then clear that the region spanned by two different transformations (say, either side of the identity) will not be convex. By contrast, two non-affine transformations (as,

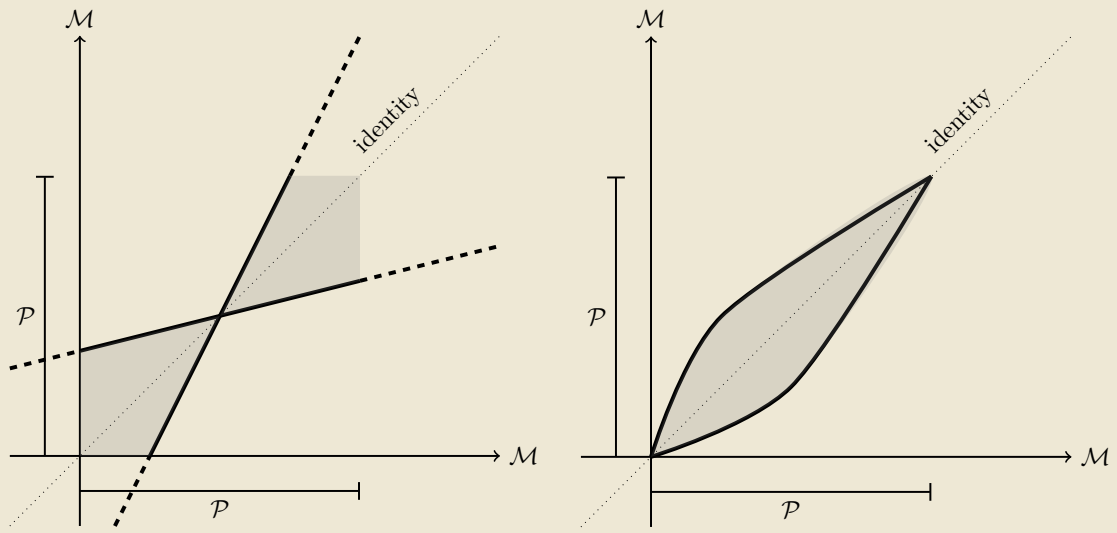


**Fig. 1.** A non-convex subset of $\mathcal{M}^2$ bounded by affine transformations, and a convex subset of $\mathcal{M}^2$ bounded by non-affine transformations.

for example, in logistic regression) either side of the identity *may* span a convex subset of $\mathcal{M}^2$. Figure 1 attempts to illustrate both these points.

Unfortunately, collapsibility is a reasonably rare property among existing statistical models, and most especially among those that are closed. We conjecture that, for finite-dimensional law spaces $\mathcal{M}$, there are no nontrivial closed, affine flows. By contrast, when $\mathcal{M}$ is infinite-dimensional, there are rich families of transformations that are closed and affine, and hence strictly collapsible: see Section 4.1.1 for one such example.

### 3.3.7. *Invariance*

We close this section by highlighting an invariance property shared by all flows, namely invariance to recoding of covariates. This is, in essence, a feature of the vector space structure of a flow $f^{\mathbb{V}}$, and says that if we recode our covariate comparisons in a suitable manner, we can reproduce the previous model fit using different parameter values. Formally, let $T\colon \mathbb{U} \to \mathbb{U}$ be an invertible transformation, for example recoding a binary contrast $X$ as $X' = XT = 1 - X$. Then for any $\theta \in \Theta$ we can reproduce the linear predictor $\eta(\theta) = X\theta$ as $\eta'(\theta') = X'\theta' = (XT)(T^{-1}\theta)$.

As a simple example, if $f^v$ maps the law $Q$ for the reference group of untreated individuals to the treated law $P$, but we choose to recode things so that treated subjects are now the reference group, we have of course that $Q = Pf^{-v}$. This invariance to recoding of covariates is a feature of the *invertibility* of each element of a flow.

We might also like to know if our flows are invariant to coding of the outcome. We can express this in terms of an invertible transformation $T\colon \Upsilon \to \Upsilon$ that recodes $Y$, and its associated transformation $g$ of $\mathcal{M}$, defined by $(Pg)(A) = P(AT^{-1})$. Standard examples are often involutions: the recoding of a binary variable $Y$ as $Y' = YT = 1 - Y$, for instance, or the recoding of a continuous variable $Y$ as $Y' = YT = -Y$. A flow is invariant to such a transformation if $gf^{\mathbb{V}}g^{-1} = f^{\mathbb{V}}$; that

is, if the spanned model spaces are equivalent irrespective of coding. Unlike recoding covariates, invariance to such recoding of outcomes is far from guaranteed, and is a stronger condition than weak commutativity of the subgroups $\langle g \rangle$ and $f^{\mathbb{V}}$. This kind of invariance, sometimes called *symmetry*, is discussed in Yates (1955) and Aranda-Ordaz (1981).

### *3.4. Model fitting*

Regression by composition features a likelihood function that depends on a finite number of unconstrained linear maps $\theta_1, \ldots, \theta_m$. As such, it is well suited to any likelihood-based mode of statistical inference. Here we describe a frequentist approach, but adaptations to suit pure-likelihood or Bayesian perspectives are of course possible.

We have argued that the modular nature of regression by composition is convenient for theoretical considerations, allowing us to assess local features of a model and, if desired, to construct variations on existing models through a simple exchange of flows. We now contend that this modularity is also useful in practical model fitting, since we can express the likelihood and its derivatives in terms of a compact library of deterministic functions. Differentiation is particularly straightforward for likelihoods built by composition, and amounts to repeated application of the chain rule.

The intermediary transformations $F_j$ map laws to laws so, in order to compute their derivatives, an abstraction of differentiation to arbitrary vector spaces is needed. The *Fréchet derivative* is ideally suited to this task (see, for example, Dieudonné, 1960, pp. 143 sqq.). As briefly mentioned earlier, we shall adopt a frugal, functional notation for the Fréchet derivative: for a function of multiple variables such as $f\colon A \times B \to C$ from sets $A$ and $B$ into $C$, partial differentiation with respect to the first, $A$-valued argument is written $f^{(A)}$, and with respect to the second, $B$-valued argument as $f^{(B)}$. The function $f^{(A)}\colon A \times B \to \mathcal{L}(A, C)$ outputs the best linear approximation to the function from $A \to C$ holding the $B$-valued argument fixed, which we identify with (and do not distinguish notationally from) the function $f^{(A)}\colon A \times B \times A \to C$ that is linear in its third argument.

In general, a regression-by-composition likelihood will not have a closed form that can be globally maximized analytically; numerical optimization will typically be needed. However, the *local* components of a regression by composition (the flows $f_j$ and local covariate contrasts $X_j$) can all have explicit expressions that are amenable to analysis. The log-likelihood contribution of a single individual can be specified, somewhat abstractly, as a (function-valued) random variable $\ell\colon \Omega \times \prod_j \Theta_j \to \mathbb{R}$ given by the composition $\ell = \mathbb{P}_m L$, where $L\colon \mathcal{M} \times \Omega \to \mathbb{R}$ is a $\mathcal{Y}$-measurable random function satisfying

$$PL = \log \frac{\mathrm{d}P}{\mathrm{d}Q}(Y)$$

and where $\mathrm{d}P/\mathrm{d}Q$ is the probability density function of $P$ or, more generally, the Radon-Nikodym derivative of $P$ with respect to some suitable reference law $Q$ (Nikodym, 1930). Specification of the function $L$ is challenging but important. For finite discrete distributions (binary outcomes, say) and appropriate counting measure $Q$, it may be possible to describe $L$ very simply in terms of the natural Euclidean representation of the law $P$ (Section 3.2.2). For censored survival outcomes, both the survivor function and its first (time) derivative play a role in the usual likelihood, so a fitting routine will either need access to both functions or appeal to some form of automatic or numerical differentiation. For continuous outcomes represented via a moment generating function or characteristic function, the function $L$ amounts to inversion of the Laplace or Fourier transforms,

perhaps by way of a saddlepoint approximation (Daniels, 1954). It is interesting to note that it is through this final component $L$, and only through $L$, that the outcome data $Y$ enters.

If exhaustive flows are employed in the early stages of a regression by composition, then choice of the initial law $\mathbb{P}_0$ is essentially arbitrary, because the first (or indeed subsequent) flows can modify it in arbitrary ways. In many cases it could be given a default value based on the range $\Upsilon$ of $Y$ (say, a standard normal law if $\Upsilon = \mathbb{R}$). The only material requirement is that $\mathbb{P}_0$ should not be a fixed point of the first flow $f_1$, perhaps achieved by ensuring that $\mathbb{P}_0$ has support across the whole of $\Upsilon$. Maximum entropy distributions (Jaynes, 1957), perhaps matching one or two empirical moments of the outcome $Y$, seem especially natural in this context. If no exhuaustive flows are employed then the choice of $\mathbb{P}_0$ *is* important, and restricts the model space to those reachable from this starting point via the specified flows.

There is also a modicum of input required from the analyst in the specification of bases for $\mathbb{U}_j$ and $\mathbb{V}_j$. These choices are also independent of the data to be analysed, and again default options can be supplied to the user (for example, the standard basis of coordinate space).

Given $\mathbb{P}_0$, evaluation of the log-likelihood function then depends principally on the flows $f_j$ and the likelihood 'converter' $L$, all of which are deterministic objects that can be stored in a computer program. In combination with a numerical routine for derivative-free function optimization, this is all that is needed to maximize the likelihood and compute the maximum likelihood estimates $\tilde{\theta}_1, \ldots, \tilde{\theta}_m$ and, more interpretably,

$$\tilde{\mathbb{P}} = \mathbb{P}_m(\tilde{\theta}_1, \ldots, \tilde{\theta}_m).$$

However, we can also compute log-likelihood derivatives relatively easily, with corresponding advantages in terms of efficient likelihood maximization and direct evaluation of the (inverse) information matrix. It turns out that the only additional ingredients needed are the partial derivative $L^{(\mathcal{M})}$ of $L$ with respect to its law-valued argument, and the two partial derivatives $f_j^{(\mathcal{M})}$ and $f_j^{(\mathbb{V}_j)}$ of the flows $f_j$ with respect to their law-valued and vector-valued arguments, respectively. To see why, recall that $\ell = \mathbb{P}_m L$. For $k = 1, \ldots, m$, we then have by the chain rule that

$$\ell^{(\Theta_k)} = \mathbb{P}_m^{(\Theta_k)} \circ (\mathbb{P}_m L^{(\mathcal{M})}),$$

where we use the $\circ$ symbol to emphasize that the right-hand side first applies the function $\mathbb{P}_m^{(\Theta_k)}$ and then *composes* it with application of the function $\mathbb{P}_m L^{(\mathcal{M})}$, itself a composition of two functions. This is where the partial derivative $L^{(\mathcal{M})}$ enters the computation.

Terms like $f_j^{(\mathcal{M})}$ emerge from the recursive relationship $\mathbb{P}_j = \mathbb{P}_{j-1} F_j$, which means that

$$\mathbb{P}_j^{(\Theta_k)} = \mathbb{P}_{j-1}^{(\Theta_k)} \circ (\mathbb{P}_{j-1} F_j^{(\mathcal{M})})$$

for $k < j \leq m$. Because $F_j = \eta_j f_j$, we may write $F_j^{(\mathcal{M})} = \eta_j f_j^{(\mathcal{M})}$; no chain rule is required here, where the argument with respect to which we are differentiating occurs in the *last* function being composed. After iterating down to $j = k$, we have

$$F_k^{(\Theta_k)} = \eta_k^{(\Theta_k)} \circ (\eta_k f_k^{(\mathbb{V}_k)}),$$

whence the need for the partial derivatives $f_j^{(\mathbb{V}_j)}$. Finally, since $\eta_k = X_k \beta_k$, it follows that

$$\eta_k^{(\Theta_k)} = X_k \beta_k^{(\Theta_k)}.$$

These last two derivatives are entirely straightforward to compute since $\eta_j$ is, by construction, linear in $\theta_j$ and, even more simply, $\beta_j$ acts on $\theta_j$ as a somewhat baroque identity function.

## 4. A medley of flows

Flows are the most important ingredients in any regression by composition. In this section we introduce a few of the many possible flavours of flow, and discuss some of their properties. Our catalogue is far from exhaustive; instead the aim is to give concrete examples of some of the ideas that we have so far presented only in a general setting. These illustrations make use of a variety of characterizations of probability measures, including the cumulative distribution function, the survivor function, and the characteristic function.

Some flows work equally well when applied to outcomes of different type (e.g. real-valued, positive-valued, count-valued), while others are more tailored to particular kinds of response (e.g. survival times). Rather than organize flows by the range of the outcome, then, we first describe four generic classes of flows before turning our attention to specific instances that do not (to our knowledge) belong to any of these classes. Somewhat surprisingly, at least to us, the connections across multiple outcome types are of a qualitatively different character to the unification offered by generalized linear models. As we shall see, generalized linear models do not extend uniquely to a single regression by composition formulation, for the simple reason that, except in the binary case†, generalized linear models do not specify how an arbitrary law should be transformed by the action of the flow.

### *4.1. Generic flows*

#### 4.1.1. *Quantile manipulation*

For laws with a continuous, ordered range $\Upsilon$, for example $\Upsilon = \mathbb{R}$ or $\mathbb{R}_+$, a large class of flows arise as transformations of the quantiles of the input law. Such flows can be most easily expressed in terms of either the cumulative distribution function or the survivor function. Given a set of strictly increasing functions $h_v : \Upsilon \to \Upsilon$ mapping old quantiles to new quantiles, one for each $v \in \mathbb{V}$, then the corresponding flow transforms the cumulative distribution function (or survivor function) $\hat{P}$ as

$$(\hat{P} f^v)(t) = \hat{P}\{h_v^{-1}(t)\}$$

for every $t \in \Upsilon$ and $v \in \mathbb{V}$. Quantile transformations of this type are discussed extensively by Hothorn et al. (2018). Examples include the location shift flow for which $h_v(t) = t + v$ and where $v \in \mathbb{R}$, the scaling (or loglinear) flow $h_v(t) = vt$ with $v \in \mathbb{R}_+$, and (for $\Upsilon = \mathbb{R}_+$) the power flow $h_v(t) = t^v$, where $v \in \mathbb{R}_+$, which is closely related to the Box-Cox transformation (Box & Cox, 1964). All flows manipulating quantiles in this way are linear (and hence affine) and closed, with fixed points at $t \mapsto 0$ and $t \mapsto 1$. Generalizations of quantile manipulations to multivariate outcomes are reasonably straightforward.

For $\Upsilon = \mathbb{R}_+$, the scaling flow corresponds to a specific accelerated failure time model (Wei, 1992). Where the loglinear model accelerates time uniformly, the more general accelerated failure time allows the function $h_v$ to scale time dynamically: the only additional stipulation is that

†Even in the binary case there is (literally) a degree of ambiguity: strictly, binary generalized linear models only specify what happens to probability laws $P \in \mathcal{P}$, and multiple extensions to the whole of $\mathcal{M}$ are possible.

$h_v(0) = 0$. This flexibility makes the general accelerated failure time model exhaustive in the sense of Section 3.3.3. Its most natural index group (the increasing functions $\mathbb{R}_+ \to \mathbb{R}_+$ starting at 0) is, however, *not* a vector space, since composing increasing functions is not commutative. The group does have an abelian subgroup, namely the scaling (or loglinear) flows. Further investigation of flows parametrized by non-abelian Lie groups seems warranted by the combination of attractive properties exhibited by the general accelerated failure time flow.

### 4.1.2. *Extrema*

Still given an ordered range $\Upsilon$, though now not necessarily continuous, we can manipulate an input cumulative distribution function or survivor function in quite a different way by multiplying it with another such function. This has the effect of making the output random variable the *minimum* or *maximum* of the input random variable and another quantity, assumed independent. Formally, let $\hat{Q}_v$ be a set of functions $\Upsilon \to [0, 1]$, indexed by $v \in \mathbb{V}$ and closed under pointwise multiplication. Then for $\hat{P}$ a cumulative distribution function or survivor function, we can define extremum-type flows as

$$(\hat{P} f^v)(t) = \hat{P}(t)\hat{Q}_v(t)$$

for all $v \in \mathbb{V}$ and $t \in \Upsilon$. There are many examples of flows of this type: on $\Upsilon = \{0, 1\}$, and when represented respectively by the cumulative distribution function and survivor function, the risk ratio and survival ratio flows are given by $\hat{Q}_v(t) = v + (1 - v)t$, for $v \in \mathbb{R}_+$. The Aalen additive hazards model (Aalen, 1989) and 'complementary' additive hazards model are similarly defined respectively in terms of the cumulative distribution function and survivor function, but the multiplying function $\hat{Q}_v$ is of an unspecified form: indeed, this flexibility is one of the main attractions of the Aalen model. These flows are linear, with a single fixed point at either $t \mapsto 0$ or $t \mapsto 1$, depending on the choice of representation.

That the survival ratio flow and risk ratio flow are *not* equivalent in the sense of Definition 3.5 is explained by the fact that neither flow is invariant to a recoding of the outcome variable $Y$.

### 4.1.3. *Convolution*

Another broad class of flows emerges when we consider laws with ranges $\Upsilon$ that are closed under addition, for example $\Upsilon = \mathbb{R}, \mathbb{R}_+$ or $\mathbb{N}$; $\Upsilon = \{0, 1\}$ is not closed in this way because $1 + 1 \notin \{0, 1\}$. Since addition of independent random variables corresponds to multiplication of characteristic functions (or moment generating functions, or probability generating functions), this kind of flow can be naturally represented in terms of these generating functions. For concreteness we work here with the characteristic function, but exact equivalents exist for moment or probability generating functions. We allow $v \in \mathbb{V}$ to index a set of functions $\hat{Q}_v : \mathbb{R} \to \mathbb{C}$ that are continuous at 0 and satisfy $0 \mapsto 1$. Under pointwise multiplication, this set $\hat{Q}_\mathbb{V}$ is assumed to form a vector subspace (isomorphic to $\mathbb{V}$) of the set of all such functions satisfying these necessary but not sufficient conditions for a function to be a characteristic function, having identity element $t \mapsto 1$ corresponding to an almost surely zero random variable. Then for $t \in \mathbb{R}$ and $v \in \mathbb{V}$,

$$(\hat{P} f^v)(t) = \hat{P}(t)\hat{Q}_v(t)$$

characterizes a flow transforming an input random variable by adding an independent random variable with characteristic function $\hat{Q}_v$. Examples again include the location shift flow given by

$\hat{Q}_v(t) = \exp(itv)$ indexed by $v \in \mathbb{R}$, but also the addition (or subtraction) of Gaussian noise where $\hat{Q}_v(t) = \exp(-vt^2/2)$ with $v \in \mathbb{R}$, which is an example of the *heat-flow* transformations of Calin (2020). All flows defined by convolution are linear (and therefore affine), and have no fixed points.

Convolutions with discrete random variables are also possible: a Poisson increment results from setting $\hat{Q}_v(t) = \exp[v\{\exp(it) - 1\}]$ for $v \in \mathbb{R}$. Adding a Poisson increment to a Poisson input results in another Poisson law, so (at least for $v \geq 0$) the Poisson increment flow is closed on the set of Poisson distributions, and hence parametrizes one possible flavour of Poisson regression model. Clearly this $\hat{Q}_v$ cannot be a valid characteristic function if $v < 0$: this equates notionally to a Poisson random variable with negative mean and variance. Nevertheless, the flow remains well-defined by understanding $\hat{Q}_v$ as a simple multiplier rather than insisting on interpreting it as a generating function. Whenever $\hat{Q}_{\mathbb{V}}$ contains functions that are not positive-definite (e.g. $v < 0$ for Gaussian noise or a Poisson increment), the corresponding flow will not be closed.

#### 4.1.4. *Compounding*

Rather than convolving an independent random variable with the input, we can instead add together independent realizations from the input distribution itself. Because multiplying characteristic functions with themselves amounts to exponentiation, the number of copies added together need not be a natural number, and in fact can itself be drawn from an arbitrary probability distribution on the positive reals. Let $\{G_v : v \in \mathbb{V}\}$ be a set of factorial moment generating functions (of the form $t \mapsto P(t^Y)$ for some law $P$ of a positive random variable $Y$), closed under composition, with identity element $t \mapsto t$ corresponding to a constant random variable $Y = 1$. Then for $v \in \mathbb{V}$ and $t \in \mathbb{R}$, a family of compounding flows is given by

$$(\hat{P}f^v)(t) = G_v\{\hat{P}(t)\}$$

where $v$ is a positive real. Here again we are thinking of $\hat{P}$ as a characteristic function, but it could instead be another generating function. Examples include the trivial $G_v(t) = t^v$ for $v \in \mathbb{R}_+$, adding together $v$ independent copies of the input random variable, but also include compounding via the factorial moment generating function of (say) a product of $v$ independent exponential distributions with unit mean. In this latter case we have $G_1(t) = (1 - \log t)^{-1}$ so that $G_2(t) = \{1 + \log(1 - \log t)\}^{-1}$, $G_{-1}(t) = \exp(1 - 1/t)$ and of course $G_0(t) = t$. (We speculate that an interpolation to the whole of $\mathbb{V} = \mathbb{R}$ exists.) In general, compounding flows are not affine, but are typically closed, and have a fixed point at the characteristic function $t \mapsto 1$.

The canonical formulation of Poisson regression describes how covariates change Poisson distributions into other Poisson distributions, but it is not entirely obvious how to generalize this notion to map laws to laws. Compounding offers one possible route: for a positive real number $v$, the transformation $f^v$ should change an input mean $\lambda$ to an output mean $\lambda v$, and so map $\hat{P}(t) = \exp[\lambda\{\exp(it) - 1)\}]$ to $\exp[\lambda v\{\exp(it) - 1)\}] = \{\hat{P}(t)\}^v$. Simple compounding with $G_v(t) = t^v$ achieves exactly this. One interesting feature of this flow is that, in addition to mapping Poisson distributions to other Poisson distributions, it also maps negative binomial distributions to other negative binomial Distributions. Naturally, this flow has the effect of increasing (or decreasing) the variance, as well as the mean.

### 4.2. Specific flows

#### 4.2.1. *Generalized linear flow*

Let $g$ be a link function for binary data as customarily defined in generalized linear models (Nelder & Wedderburn, 1972) and, for $v \in \mathbb{R}$, let $h^v$ be the unique (real) analytic continuation of the function $t \mapsto g^{-1}\{g(t)+v\}$ to the whole real line, which Daniel et al. (2021) call the *characteristic collapsibility function*; its unconstrain-shift-reconstrain structure dates back at least to Abel (1826). Representing a law $P$ on $\{0,1\}$ through its total mass $P(\Omega)$ and the mass $P(Y=1)$ it assigns to the event $\{Y=1\}$ as

$$\hat{P} = (\hat{P}_\Omega, \hat{P}_1) = (P(\Omega), P(Y=1)),$$

define the binary generalized linear flow as

$$\hat{P} f^v = (\hat{P}_\Omega, h^v(\hat{P}_1)).$$

The properties of binary generalized linear flows follow directly from properties of $h^v$. Specifically, they are closed if $h^v$ is closed on $[0,1]$, affine if $h^v$ is affine, and have fixed points where $h^v$ has fixed points. Special cases include the linear probability flow, in which $h^v : t \mapsto t+v$ is affine (but not linear), is not closed, and has no fixed points, and the logistic regression flow, in which $h^v : t \mapsto \operatorname{expit}\{\operatorname{logit}(t)+v\} = t\exp(v)\{1-t+t\exp(v)\}^{-1}$ is not affine, is closed, and has two fixed points, at 0 and 1. A more unusual choice, with variance stabilizing properties, is the angular or arcsine-square-root link function (Cox & Snell, 1989, p. 21; Rücker et al., 2009), which leads to a flow that is closed but not affine, and has no fixed points. The arcsine-square-root link function is also notable because it is *periodic* in $v$ with period $\pi$, and therefore the identification between $\mathbb{R}$ and $f^{\mathbb{R}}$ is *not* bijective. We touch again on this point in Section 6.

The possibility of associating different generalized linear model link functions to different sets of covariates within the same regression by composition is reminiscent of the composite link functions of Thompson and Baker (1981). However, Thompson and Baker effectively allow different *subjects* or *experimental units* to have different link functions, while regression by composition allows different *covariates* or *sets of covariates* to have different link functions.

#### 4.2.2. *Hybrid relative risk*

The risk ratio flow gives rise to all transformations from $P(Y=1)$ to $(Pf^v)(Y=1)$ that, on a l'Abbé plot, can be viewed as straight lines through $(0,0)$ with positive gradient. The survival ratio flow similarly gives rise to all straight lines through $(1,1)$ with positive gradient. By combining half of one set and half of the other, the set of transformations corresponding to the switch relative risk (van der Laan et al., 2007) arises: all straight lines through $(0,0)$ or $(1,1)$ with positive gradient less than or equal to 1. While this family is not a group (since only the identity transformation has an inverse), we can extend it to a group of transformations by *composing* a risk ratio flow $f$ with multiplicative index $v$ and a survival ratio flow $g$ with multiplicative index $w$. We refer to this pair of flows as the hybrid relative risk: it encompasses *all* straight lines of any positive gradient, and has no fixed points:

$$\hat{P} f^v g^w = (\hat{P}_\Omega, 1-(1-\hat{P}_1 v)w)$$

All transformations in this group are affine, and are represented on a l'Abbé plot with intercept $1-w$ and slope $vw$. Unlike the switch relative risk model, the hybrid relative risk transformations are not all closed. The composition could be done in the opposite order, leading to the same set of

transformations: the risk ratio and survival ratio flows commute in the weak sense. We illustrate the use of the the hybrid relative risk in Section 5. Analogous hybrid versions of the Aalen and complementary Aalen flows are also possible, removing even the one fixed point that these flows have in isolation.

#### 4.2.3. *Cox flow*

Although the representation $\hat{P}$ of a survival distribution $P$ can be many things, one thing it cannot be is a hazard function. To appreciate why, imagine trying to define an isomorphism $\psi\colon \mathcal{M} \to \tilde{\mathcal{M}}$ that takes a law $P$ to its associated hazard function $\tilde{P}$, given by

$$\tilde{P}(t) = -\frac{\hat{P}'(t)}{\hat{P}(t)},$$

where $\hat{P}$ is the survivor function of $P$, and $\hat{P}'$ its derivative. While $\hat{P}$ and $\hat{P}'$ are related linearly to $P$, the hazard function $\tilde{P}$ is visibly not so, and hence $\psi$ cannot be an isomorphism. As Sjölander et al. (2016) explain, it is conditioning on past survival (division by $\hat{P}(t)$) that causes problems here, making $\psi$ nonlinear. The *cumulative* hazard function does *not* share this defect: see Section 4.1.2.

Therefore, although the Cox model (Cox, 1972) is most usually expressed in terms of hazard functions, we here describe how the Cox flow transforms a survivor function $\hat{P}$ as

$$(\hat{P} f^v)(t) = \prod_{s=0}^{t} \{\mathrm{d}\hat{P}(s)\}^{v(s)}$$

where again $\mathrm{d}\hat{P}(s)$ is a multiplicative increment, and $v$ is a positive-valued function $[0, \infty) \to (0, \infty)$. The elements of the group $\mathbb{V}$ of positive-valued functions here combine via pointwise multiplication. The Cox flow is closed but not affine, and has two fixed points, at $t \mapsto 0$ and $t \mapsto 1$.

#### 4.2.4. *Hyperscaling*

The compounding flow is not the only one that is closed on the set of Poisson laws. Representing a law $P$ as its probability generating function $\hat{P}$, Laurent (2012) introduces a hyperscaling operation defined as

$$(\hat{P} f^v)(t) = \hat{P}\{1 - v(1 - t)\},$$

which also maps a Poisson distribution with mean $\lambda$ to a Poisson distribution with mean $\lambda v$. The hyperscaling flow is not closed on the negative binomial distributions, however. Like the mutliplicative count flow, hyperscaling has a fixed point at $\hat{P} = t \mapsto 1$. Further, it is also an affine flow. This therefore seems to offer an attractive, collapsible alternative to 'standard' Poisson regression.

A multivariate version of hyperscaling exists: writing $\hat{P}$ for the probability generating function $\hat{P}\colon (t_1, \ldots, t_k) \mapsto P(t_1^{Y_1} \cdots t_k^{Y_k})$, then for $v = (v_1, \ldots, v_k) \in \mathbb{R}^k$ we have

$$(\hat{P} f^v)(t_1, \ldots, t_k) = \hat{P}\{1 - v_1(1 - t_1), \ldots, 1 - v_k(1 - t_k)\}.$$

This multivariate flow has the effect of multiplying each mean by the corresponding $v_j$. This could be used as a much more direct alternative to the somewhat complicated Dirichlet negative multinomial regression proposed by Farewell and Farewell (2013).

#### 4.2.5. *Zero-inflation flow*

In their usual formulation (see, e.g., Lambert, 1992), zero-inflated count regressions require that we imagine a latent binary regression identifying 'excess' zeros, and a latent count regression giving the count-valued outcome conditional on the observation *not* being an excess zero. Both sets of coefficients are identified only by relying on distributional assumptions; if the distribution of the latent count changes, then the implied proportion of excess zeros is different, a fact that could not be diagnosed from the observed data.

An attempt to write a standard zero-inflated model as a regression by composition would fail because transformations would need to be applied to the laws of the latent binary and count variables, as opposed to the law of $Y$. We propose instead that a zero-inflated regression by composition model be formulated by incorporating a zero-inflation flow, which alters $P(Y = 0)$ and scales all other probabilities $P(Y = y)$ for $y \geq 1$ by a normalizing constant.

Formally, the zero-inflation flow could be defined with a multiplicative index $v \in \mathbb{R}_+$ as

$$(\hat{P} f^v)(t) = \left\{ \frac{\hat{P}(t) - \hat{P}(0)}{1 - \hat{P}(0)} \right\} \{1 - \hat{P}(0)v\} + \hat{P}(0)v.$$

This zero-inflation flow deploys a risk ratio flow to change the probability $P(Y = 0) = \hat{P}(0)$; clearly many other choices are possible.

A zero-inflated regression by composition can then be formed by composing (say) a Poisson flow and the zero-inflation flow—another example of hybrid flows—for each covariate contrast in turn. An advantage of this approach over the traditional formulation is that it does not rely on latent variables, and that inferences relying wholly on unverifiable distributional assumptions are not invited. Another advantage is that the regression-by-composition formulation is not confined to a particular choice of input law $\mathbb{P}_0$, meaning that Poisson, negative binomial, Poisson-inverse-Gamma, or any other distribution can be inserted with ease. Alternative choices of flows are of course also possible, including flexible affine choices such as the hyperscaling flow (Section 4.2.4) composed with two zero-inflation flows, based on the generalized linear transformation with a log and complementary log link, respectively.

## 5. Worked data examples

We undertake in this section reanalyses of two freely-available study datasets. Our aims are, firstly, to allow others to access the data and so to explore for themselves the substantive features we claim regression by composition helps to tease out and, secondly, to illustrate as clearly as possible some of the more unconventional flexibility of regression by composition. In addition, and especially for the first worked example, we use this as an opportunity to exemplify some of the less familiar notational choices made in the earlier sections of the paper.

### *5.1. Real-valued outcome: the Land Rent data*

Our first example is a reanalysis of the Land Rent Data (Weisberg, 2005, p. 208), arising from a 1977 investigation of land rent prices for 67 counties in Minnesota. The outcome $Y$ is the county-wide average rental cost (in United States dollars) per acre of farmland used to grow alfalfa. In this analysis, we include two of the available covariates. The first ($Q$) is the average rent (in dollars) paid for all tillable land in the county, acting as a proxy for soil quality. The second

($L \in \{\text{Yes}, \text{No}\}$) is whether liming is required in the county: liming reduces soil acidity, but the associated costs are thought to depress the rental price for land used to grow alfalfa.

The support of the law of $Y$ is positive (although see the comment at the end of Section 5.1.6), and we may conjecture that the dependence of the conditional law of $Y$ on $Q$ is multiplicative, for example in the sense that as the soil quality of land that does not require liming doubles, the quantiles of the law of the rental cost increase $\theta$-fold, for some (positive) $\theta$. Using conventional methods, this may lead us to consider taking the logarithm of both the outcome $Y$ and the covariate $Q$. The dependence of the conditional law of $Y$ on $L$, on the other hand, may be closer to additive (assuming that there is a fixed per-acre cost of liming). Using regression by composition, combining such multiplicative and additive relationships for different covariates is possible.

### 5.1.1. *Flows*

We use the location shift and scale flows introduced previously, i.e.

$$(\hat{P} f^v_{\text{sh}})(t) = \hat{P}(t - v),\ v \in (\mathbb{R}, +)$$

and

$$(\hat{P} f^v_{\text{sc}})(t) = \hat{P}(t/v),\ v \in (\mathbb{R}_+, \times)$$

where $\hat{P}$ is a cumulative distribution function. We also make use of a new ('power') flow:

$$(\hat{P} f^v_{\text{po}})(t) = \hat{P}(t^{1/v}),\ v \in (\mathbb{R}_+, \times).$$

Using the notation of Section 3.2.3, when $\mathbb{V} = (\mathbb{R}, +)$ its basis is $(v^1) = (1)$, and when $\mathbb{V} = (\mathbb{R}_+, \times)$ its basis is $(v^1) = (e)$ (Euler's number). Note that $q$ is always 1 in this example: our flows have one-dimensional indices.

These flows correspond to different types of dependencies of the conditional law of $Y$ on covariate contrasts. For example, the location shift flow encodes an additive dependence, whereas the scale flow encodes a multiplicative dependence. There is also an obvious connection between the three flows and additive, multiplicative and power transformations of the outcome variable $Y$ itself. The power flow is closely related to the Box–Cox transformation (Box & Cox, 1964), although we do not consider negative $v$ here. An example of each flow type is illustrated in Figure 2.

### 5.1.2. *Covariate embedding, filtration and ordering*

Using the notation of Section 3.2.3, we choose to encode $Q$ as a multiplicative contrast (relative to unity) taking values in $\mathbb{U} = (\mathbb{R}_+, \times)$ with the single basis vector $(u^1) = (e)$. On the other hand, we choose to encode $L$ as an additive contrast (relative to 'No') which we can take to be in $\mathbb{U} = (\mathbb{R}, +)$ with the single basis vector $(u^1) = (1)$.

For the multiplicative contrasts derived from $Q$, each value $X = Q$ is associated with a basis vector coefficient $X_1 \in \mathbb{R}$, so that $X_1$ 'scalar multiplied' by $u^1$ in the vector space $(\mathbb{R}_+, \times)$ is equal to $X$. This means that $X = e^{X_1}$, or $X_1 = \log(X)$. In this sense, considering multiplicative contrasts of $Q$ is equivalent to taking logs. For the binary covariate $X = \mathbb{1}\{L = \text{Yes}\}$, the covariate value $X$ (either 1 or 0) and its basis vector coefficient $X_1$, are the same.

The covariate filtration we consider for our first three models is $\mathcal{F}_1 = \mathcal{F}_2 = \sigma(\emptyset)$, $\mathcal{F}_3 = \sigma(Q)$, $\mathcal{F}_4 = \sigma(L)$. This corresponds to an ordering of the embedded covariate contrasts as follows:

$X_1 = X_2 = 1$, $X_3 = Q$, $X_4 = \mathbb{1}\{L = \text{Yes}\}$. Starting with $\mathcal{F}_1 = \mathcal{F}_2$ allows us the flexibility to change $\mathbb{P}_0$ in two ways (e.g. with both a shift and scale flow, or both a scale and power flow). Note that each $p$ (the dimension of $\mathbb{U}$) is 1 in these three models.

For reasons discussed below, in our fourth model, we consider a different filtration, namely $\mathcal{F}_1 = \mathcal{F}_2 = \sigma(\emptyset)$, $\mathcal{F}_3 = \mathcal{F}_4 = \sigma(L)$, $\mathcal{F}_5 = \mathcal{F}_6 = \sigma(L, Q)$, and a correspondingly different ordering of the embedded covariate contrasts as follows: $\tilde{X}_1 = \tilde{X}_2 = 1$, $\tilde{X}_3 = \tilde{X}_4 = \mathbb{1}\{L = \text{Yes}\}$, $\tilde{X}_5 = \tilde{X}_6 = \left[Q^{\mathbb{1}\{L=\text{No}\}}, Q^{\mathbb{1}\{L=\text{Yes}\}}\right]$. The two-dimensional vector space $\mathbb{U}$ in which $\tilde{X}_5$ and $\tilde{X}_6$ are situated is $(\mathbb{R}^2_+, \times)$ where $\times$ denotes the Hadamard product (element-wise multiplication) with basis vectors $(u^1, u^2) = \left(\begin{pmatrix} e \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ e \end{pmatrix}\right)$. The basis vector coefficients $\tilde{X}_{5,1}$ and $\tilde{X}_{5,2}$ corresponding to $\tilde{X}_5$ are respectively $\mathbb{1}\{L = \text{No}\} \log Q$ and $\mathbb{1}\{L = \text{Yes}\} \log Q$, which follows from

$$\tilde{X}_5 = \tilde{X}_{5,1} \odot \begin{pmatrix} e \\ 1 \end{pmatrix} \oplus \tilde{X}_{5,2} \odot \begin{pmatrix} 1 \\ e \end{pmatrix} = \begin{pmatrix} \exp(\tilde{X}_{5,1}) \\ \exp(\tilde{X}_{5,2}) \end{pmatrix}$$

where $\oplus$ and $\odot$ denote addition and scalar multiplication in $(\mathbb{R}^2_+, \times)$.

### 5.1.3. *A note on additive and multiplicative dependencies and contrasts*

The notion of an additive or multiplicative dependence, achieved by employing a shift or a scale flow, respectively, of a conditional outcome law on a covariate, should be separated from the notion of an additive or multiplicative covariate contrast. The former is akin to transformation (or not) of the outcome while the latter is akin to transformation (or not) of the covariate. A flexibility afforded by regression by composition is that the effective transformation (or not) of the outcome need not remain constant across all components of the model, and that furthermore—by using multiple different flows for the same covariate contrast—dependencies of a more complex nature can be constructed.

### 5.1.4. *Linear maps and their matrix representations*

Each of the linear maps $\theta_j : \mathbb{U}_j \to \mathbb{V}_j$ ($j = 1, \ldots, m$, where $m = 4$ in the first three models, and $m = 6$ in the final model) is associated with its $(p_j \times 1)$ matrix $[\theta_j]$. For a value $x_j$ of the covariate contrast $X_j$ with associated basis vector coefficients $x_{j,k}$, we have:

$$x_j \theta_j = \underbrace{(x_{j,k} u_j^k)\theta_j}_{\text{(a)}} = \underbrace{x_{j,k}(u_j^k \theta_j)}_{\text{(b)}} = \underbrace{x_{j,k}[\theta_j]_1^k v_j^1}_{\text{(c)}}$$

where addition and scalar multiplication in (a) are as defined in $\mathbb{U}_j$, and in (b) and (c) are as defined in $\mathbb{V}_j$.

In the first model, as we will explain in Section 5.1.5, $\mathbb{V}_j = (\mathbb{R}_+, \times)$ for all $j$. This leads to:

$$\begin{aligned} X_1\theta_1 &= \exp\left([\theta_1]_1^1\right) \\ X_2\theta_2 &= \exp\left([\theta_2]_1^1\right) \\ X_3\theta_3 &= \exp\left([\theta_3]_1^1 X_{3,1}\right) = \exp\left([\theta_3]_1^1 \log Q\right) = Q^{[\theta_3]_1^1} \\ X_4\theta_4 &= \exp\left([\theta_4]_1^1 X_{4,1}\right) = \exp\left([\theta_4]_1^1 \mathbb{1}\{L = \text{Yes}\}\right) \end{aligned} \tag{1}$$

The linear maps in Models 2 and 3 are the same, except that $\mathbb{V}_4 = (\mathbb{R}, +)$, leading to:

$$X_4\theta_4 = [\theta_4]_1^1 X_{4,1} = [\theta_4]_1^1 \mathbb{1}\{L = \text{Yes}\} \tag{2}$$

In the final model, as we will see below, $\mathbb{V}_1 = \mathbb{V}_2 = \mathbb{V}_3 = \mathbb{V}_5 = (\mathbb{R}_+, \times)$ and $\mathbb{V}_4 = \mathbb{V}_6 = (\mathbb{R}, +)$. To avoid confusion we use a $\tilde{\cdot}$ on both the $X_j$ and $\theta_j$. The linear maps are then:

$$\begin{aligned} \tilde{X}_1\tilde{\theta}_1 &= \exp\left([\tilde{\theta}_1]_1^1\right) \\ \tilde{X}_2\tilde{\theta}_2 &= \exp\left([\tilde{\theta}_2]_1^1\right) \\ \tilde{X}_3\tilde{\theta}_3 &= \exp\left([\tilde{\theta}_3]_1^1 \tilde{X}_{3,1}\right) = \exp\left([\tilde{\theta}_3]_1^1 \mathbb{1}\{L = \text{Yes}\}\right) \\ \tilde{X}_4\tilde{\theta}_4 &= [\tilde{\theta}_4]_1^1 X_{4,1} = [\tilde{\theta}_4]_1^1 \mathbb{1}\{L = \text{Yes}\} \\ \tilde{X}_5\tilde{\theta}_5 &= \exp\left([\tilde{\theta}_5]_1^1 X_{5,1} + [\tilde{\theta}_5]_1^2 X_{5,2}\right) \\ &= \exp\left([\tilde{\theta}_5]_1^1 \mathbb{1}\{L = \text{No}\}\log(Q) + [\tilde{\theta}_5]_1^2 \mathbb{1}\{L = \text{Yes}\}\log(Q)\right) \\ &= Q^{[\tilde{\theta}_5]_1^1 \mathbb{1}\{L=\text{No}\} + [\tilde{\theta}_5]_1^2 \mathbb{1}\{L=\text{Yes}\}} \\ \tilde{X}_6\tilde{\theta}_6 &= [\tilde{\theta}_6]_1^1 X_{6,1} + [\tilde{\theta}_6]_1^2 X_{6,2} = [\tilde{\theta}_6]_1^1 \mathbb{1}\{L = \text{No}\}\log(Q) + [\tilde{\theta}_6]_1^2 \mathbb{1}\{L = \text{Yes}\}\log(Q). \end{aligned} \tag{3}$$

5.1.5. *Model* 1: *generalized linear model with log-transformed Y and Q*

The first regression by composition model we consider corresponds exactly to a generalized linear model with Gaussian errors where both $Y$ and $Q$ are log-transformed. This is done in the regression by composition framework without transforming $Y$. Instead, we let $\mathbb{P}_0$ be the standard log-normal distribution and first apply both the power and scale flows as the 'intercept' flows, that is the power flow for $X_1$ followed by the scale flow for $X_2$. These are followed by a scale flow for the multiplicative contrast for soil quality, $X_3$, and finally, a scale flow for the additive contrast for liming, $X_4$. This is equivalent (in the sense of giving numerically identical maximum likelihood estimates) to first log-transforming $Y$, starting from a standard normal $\mathbb{P}_0$, applying both scale and shift 'intercept' flows, followed by a shift flow each for $\log(Q)$ and $\mathbb{1}\{L = \text{Yes}\}$: a generalized linear model for $\log(Y)$ given $\log(Q)$ and $\mathbb{1}\{L = \text{Yes}\}$ with Gaussian errors (the variance of which are estimated in the first scale flow).

As described in Section 3.2.3, we write $\beta_j$ for the function $\mathbb{U}_j \times \Theta_j \to \mathbb{V}_j$ that maps $(X_j, \theta_j)$ to $X_j\theta_j$, and finally $\eta_j$ (the local linear predictor) for the function $\Omega \times \Theta_j \to \mathbb{V}_j$ that maps $(\omega, \theta_j)$ to $(X_j(\omega))\theta_j$. For a particular $\omega$ and the set of linear maps $\{\theta_j : j = 1, \ldots, 4\}$ defined in (1),

the conditional cumulative distribution function $\widehat{\mathbb{P}}_4$ of the outcome is written as a composition of transformations applied to $\widehat{\mathbb{P}}_0$:

$$\widehat{\mathbb{P}}_4 = \widehat{\mathbb{P}}_0 F_1 F_2 F_3 F_4 = \widehat{\mathbb{P}}_0(\eta_1 f_{\text{po}})(\eta_2 f_{\text{sc}})(\eta_3 f_{\text{sc}})(\eta_4 f_{\text{sc}}) \tag{4}$$

with $\widehat{\mathbb{P}}_0$ the cumulative distribution function of a standard log-normal distribution.

#### 5.1.6. *Model* 2: *an additive dependence on liming*

The second regression by composition model differs from the above simply by changing the final flow from a scale to a location shift flow, thus changing the nature of the posited conditional dependence of $Y$ on the liming contrast to be additive.

$$\widehat{\mathbb{P}}_4 = \widehat{\mathbb{P}}_0 F_1 F_2 F_3 F_4 = \widehat{\mathbb{P}}_0(\eta_1 f_{\text{po}})(\eta_2 f_{\text{sc}})(\eta_3 f_{\text{sc}})(\eta_4 f_{\text{sh}}) \tag{5}$$

with $\widehat{\mathbb{P}}_0$ again the cumulative distribution function of a standard log-normal distribution.

Note that the inclusion of a location shift flow (in contrast to model 1) at the end could result in the support of $Y$ including negative values in counties where liming is required. We see this as a desirable feature; supposing that (where liming is not required) only a very low rent could be charged for land with very low soil quality, then this rent would effectively be negative if an additional outlay for liming were required.

#### 5.1.7. *Model* 3: *gamma distribution*

To further emphasize the modular nature of regression by composition, the third model differs from the second simply by changing $\widehat{\mathbb{P}}_0$ to be the cumulative distribution function of an exponential distribution with unit mean.

#### 5.1.8. *Model* 4: *more flexible*

In our final regression by composition model, we return to the log-normal $\mathbb{P}_0$ but include both a shift and a scale flow for both the multiplicative soil quality contrast and the additive liming contrast, and we allow both the shift and scale parameters for soil quality to differ according to the level of liming. This is most naturally achieved by reversing the order of liming and soil quality in the model, as shown in (3). The model is then:

$$\widehat{\mathbb{P}}_6 = \widehat{\mathbb{P}}_0 F_1 F_2 F_3 F_4 F_5 F_6 = \widehat{\mathbb{P}}_0(\eta_1 f_{\text{po}})(\eta_2 f_{\text{sc}})(\eta_3 f_{\text{sc}})(\eta_4 f_{\text{sh}})(\eta_5 f_{\text{sc}})(\eta_6 f_{\text{sh}}) \tag{6}$$

with $\widehat{\mathbb{P}}_0$ the cumulative distribution function of a standard log-normal distribution.

#### 5.1.9. *Model fitting and results*

To fit the models listed above, first the numerical derivative of the cumulative distribution function in (4)-(6) is calculated using Richardson extrapolation (Richardson, 1997) via the `numDeriv` package in R, which gives the log-likelihood contribution for each county at any $\{\theta_j : j = 1, \ldots, m\}$. The maximum likelihood estimators of each $[\theta_j]$ are then calculated using the Nelder–Mead simplex method (Nelder & Mead, 1965) implemented in the `optim` function. The standard errors of

these are estimated from the numerically differentiated Hessian matrix. These estimates and their estimated standard errors are shown in Table 1.

The implied conditional median and 95% range of the predictive distribution for $Y$ at each $(L, Q)$ for each fitted model is plotted in Figure 3. Table 2 shows a comparison of model fit using the Akaike and Bayesian information criteria. The fitted models can also be visualized by looking at how the estimated outcome densities are transformed for chosen values of the covariates (see Figure 4).

#### 5.1.10. *Remarks*

This example demonstrates many features of the regression by composition framework. Firstly, it highlights (e.g. in Figure 4) that the model considers transformations (arising from covariate contrasts) of the entire conditional outcome law, rather than, say, its first moment. In particular, and unlike in generalized linear models, the estimation of variances is brought into the same modelling framework as the estimation of mean dependence on covariates. Secondly, by using different flows for different covariates, it was possible (e.g. in our second model) to combine a multiplicative dependence on one covariate with an additive dependence on another. Thirdly, by including more than one flow for the same covariate, a less restrictive model is obtained, such as one that commits neither to a purely additive nor purely multiplicative dependence on a covariate, but contains both as special cases. Finally, we saw in the third model a further example of the modular nature of the framework, by easily changing the initial law from a standard log-normal to an exponential distribution.

From Table 2, and from the similarity of parameter estimates between Model 1 and 4 in Table 1, we see that there is little reason to prefer any of the fitted models over model 1, which is equivalent to a generalized linear model with an identity link after log-transforming both $Y$ and $Q$. The fact that there are no points in the bottom left of the scatter plots (see Figure 3) from the counties in which liming is required, is however consistent with our conjectured additive dependence on the liming requirement: when the agreed cost would be very low or even negative, the land is not rented. The fact that model 1 fits best suggests that the cost of liming may be higher for fields where the soil quality is higher, or perhaps that the need for liming is associated with another relevant feature not included in this analysis, and that the dependence of rental price on that feature is multiplicative. Had the outlay for liming been greater, we might have seen a clearer difference between the fitted models.

### *5.2. Binary outcome: Randomized controlled trial of synbiotics to prevent infant sepsis*

Our second example is from a randomized controlled trial run in rural India comparing oral synbiotics against placebo in 4,556 newborn infants (Panigrahi et al., 2017). The published analysis did not make use of the collected baseline covariates, and estimated the effect of randomization on the binary primary outcome—death or sepsis within 60 days—as a risk ratio of 0.60 (95% confidence interval 0.48–0.74), demonstrating a clear protective effect of synbiotics.

**Table 1.**
Estimated model parameters and their estimated standard errors.

| | | | Model | | | | |
|---|---|---|---|---|---|---|---|
| Covariate | Flow | Parameter | 1 (equiv. to GLM) | 2 (additive liming) | 3 (gamma distribution) | 4 (flexible) | |
| 'Intercept' | Power | $[\theta_1]_1^1$ | −1.45 (0.086) | −1.53 (0.089) | −1.54 (0.099) | | −1.38 (0.182) |
| | Scale | $[\theta_2]_1^1$ | 0.084 (0.179) | 0.227 (0.169) | 0.690 (0.191) | | −0.09 (0.575) |
| Quality | Scale | $[\theta_3]_1^1$ | 0.992 (0.050) | 0.946 (0.045) | 0.858 (0.050) | $L$ = No, $[\theta_5]_1^1$ | 1.02 (0.129) |
| | | | | | | $L$ = Yes, $[\theta_5]_1^2$ | 1.00 (0.307) |
| | Shift | | | | | $L$ = No, $[\theta_6]_1^1$ | 0.546 (1.07) |
| | | | | | | $L$ = Yes, $[\theta_6]_1^2$ | −0.900 (3.94) |
| Liming | Scale | $[\theta_4]_1^1$ | −0.221 (0.058) | | | $[\theta_3]_1^1$ | −0.159 (1.26) |
| | Shift | $[\theta_4]_1^1$ | | −6.20 (1.83) | −8.73 (2.40) | $[\theta_4]_1^1$ | 0.148 (0.520) |

**Table 2.**
Model comparison.

| Model | No. parameters | log-likelihood | AIC | BIC |
|---|---|---|---|---|
| 1 (equiv. to GLM) | 4 | −237.28 | 482.56 | 491.38 |
| 2 (additive liming) | 4 | −238.07 | 484.13 | 492.95 |
| 3 (gamma distribution) | 4 | −245.36 | 498.73 | 507.54 |
| 4 (flexible) | 8 | −237.25 | 490.49 | 508.13 |

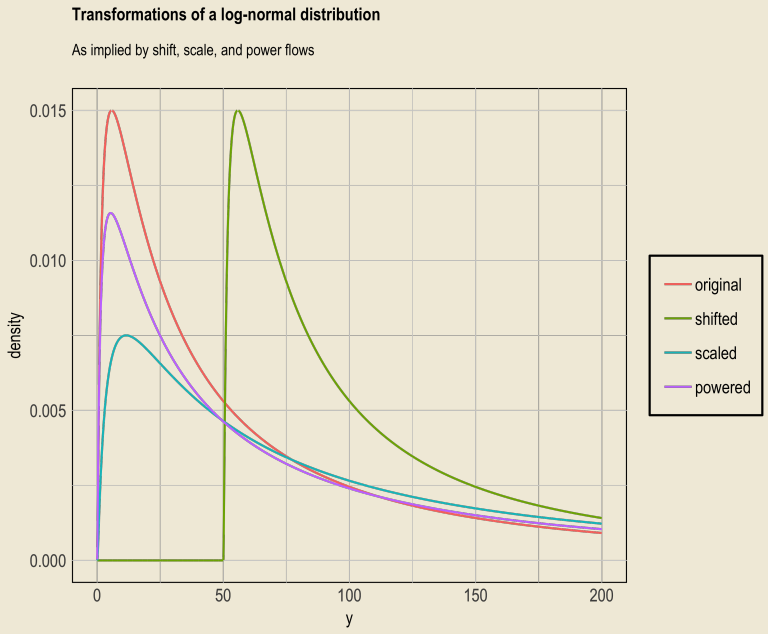


**Fig. 2.** An illustration of shift, scale and power transformations of a log-normal distribution.

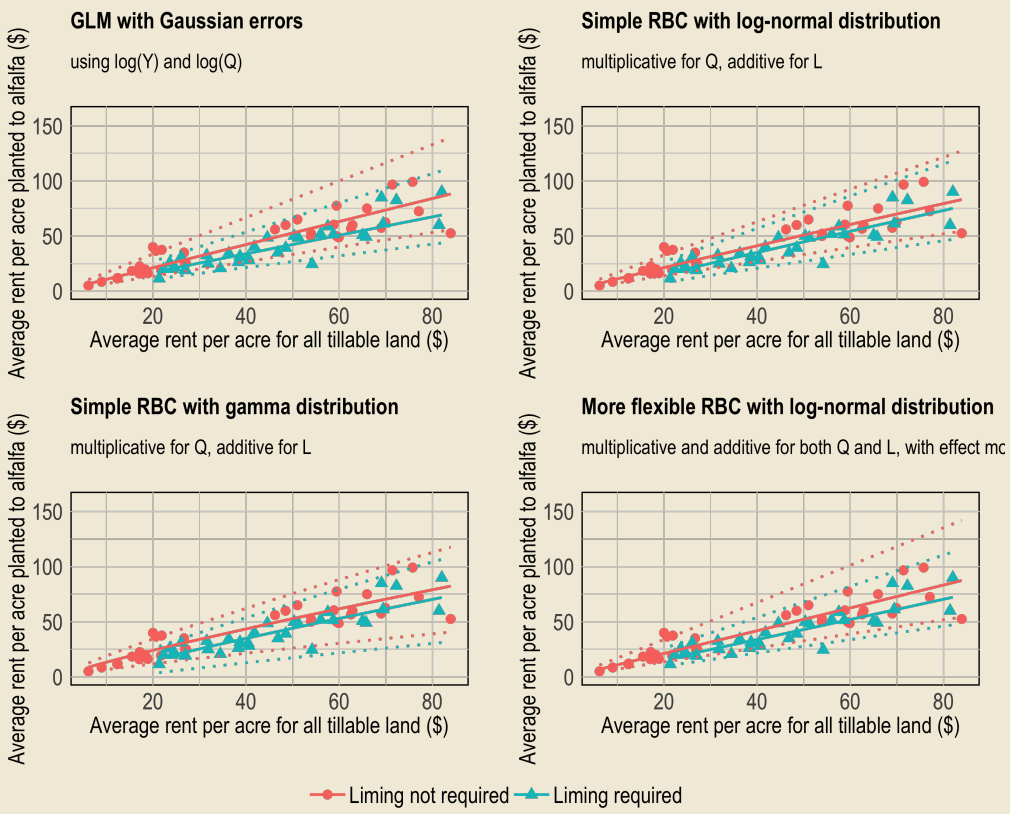


**Fig. 3.** A scatter plot showing the outcome plotted against the quantitative covariate, separately for the levels of the binary covariate. The solid lines show the pointwise median of the conditional outcome laws implied by the four fitted regression by composition models, and the dotted lines depict their corresponding 2.5th and 97.5th centiles, giving pointwise 95% predicted ranges.

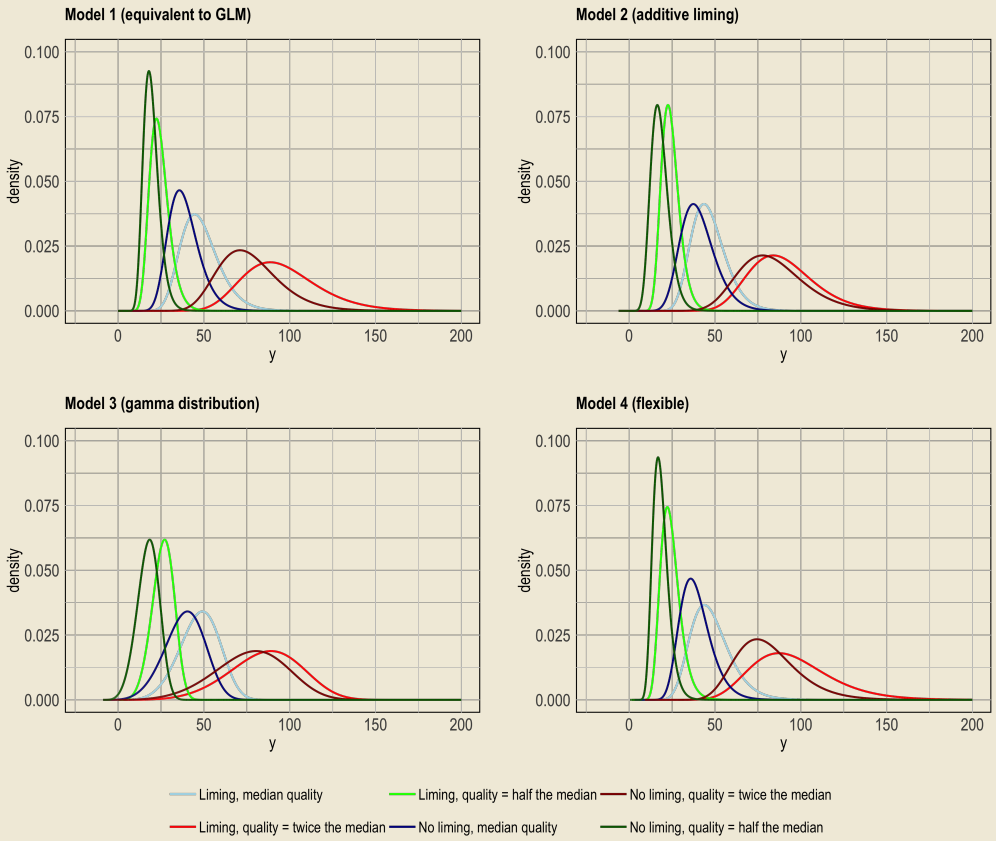


**Fig. 4.** An illustration of selected estimated conditional outcome densities as implied by the four fitted regression by composition models. The estimated conditional outcome density evaluated at the sample median value of quality ($Q$) is shown in blue, at twice this sample median is shown in red, and at half the sample median is shown in green. The lighter version of each colour is used when the conditional density is evaluated for 'liming is required', and the darker version for 'liming not required'.

Baseline covariates, including maternal age, maternal parity (1,2,3,4+), the sex of the infant, whether the infant was born prior to 37 weeks gestation, and birthweight, are available for 4,539 (99.6%) of the participants. We use these complete records to fit several possible models that take the baseline covariates into account. Using regression by composition, we can directly estimate a risk ratio (and its standard error) for the treatment effect after conditioning on baseline covariates without having to use a log link for the covariates. We also demonstrate how a hybrid relative risk for the treatment effect could instead be fitted.

### 5.2.1. *Model specifications*

We fit five models to these data. Maternal age and birthweight are used in their standardized form (i.e. after subtracting their sample mean and dividing by their sample standard deviation) throughout, and the covariate information on maternal parity is embedded into three dummy indicator variables (respectively 1 if parity is 2, 3 or $\geq 4$, and 0 otherwise) as is typically done for categorical covariates in regression models.

The first (Model 1) is equivalent to standard logistic regression but fitted as a regression by composition model. That is, a generalized linear flow with a logit link (Section 4.2.1) is selected for each covariate, with initial law $\hat{P}_0 = (1, 0.5)$, so that the model is identical to a generalized linear model with a logit link and no interactions, parameterized in terms of a baseline log odds and log odds ratios for a unit additive contrast in each embedded covariate. Recall that, as in Section 4.2.1, the representation $\hat{P}$ of a binary outcome law has $P(\Omega)$ as its first component, and $P(Y = 1)$ as its second.

Models 2 and 3 are the same as Model 1, except that they use a generalized linear flow with a log and identity link, respectively, for the treatment contrast (retaining the logit link for all other covariates), corresponding to a risk ratio and risk difference flow for treatment. Note that these are different from a log-binomial and a linear probability model, respectively, since these would use the log and identity link for all covariates. Thus these two models suffer from non-closure (predicted probabilities for new patients may not lie inside $[0, 1]$ if their estimated risk conditional on baseline covariates and placebo is outside the range of such estimated risks seen in the dataset) but only for their final flows. On the other hand, these two models directly encode treatment parameters that are collapsible, unlike Model 1, and so we need not be concerned about the baseline covariates conditioned upon when interpreting these estimates.

Model 4 is selected to illustrate the flexibility of this framework by selecting a different flow for each covariate, in the following order: a generalized linear flow with a logit link for the 'intercept' flow, a log link (risk ratio flow) for maternal age, a complementary log link (survival ratio flow) for maternal parity, a probit link for infant sex, a complementary log–log link for gestational age, a Cauchit link for birthweight, and an arcsine-square-root link for treatment. Such a model is unlikely to be selected in practice, but we include it to demonstrate the potential for such modular flexibility, and how the results for such a model may best be reported.

Finally, Model 5 is like Model 1 except that it uses the composition of a risk ratio and survival ratio flow for treatment, that is the *hybrid relative risk* introduced in Section 4.2.2. The comments made above about Models 2 and 3 regarding non-closure and collapsibility apply equally here. In addition, the hybrid relative risk admits a much larger class of (straight-line) relationships between the outcome risk under placebo and active, than Models 2 and 3.

### 5.2.2. *Fitting and Results*

Table 5.2.3 gives the maximum likelihood estimates for the parameters of each of the five models, obtained using an implementation of the conjugate gradients method (Fletcher & Reeves, 1964) via the `optim` function in `R`, where all gradients were obtained by Fréchet differentiation (see Section 3.4). The standard errors are derived from the estimated Hessian matrix by the same function.

Figures 5 and 6 use Model 4 as an example and show how the composition of flows in the model transforms the conditional outcome law from the arbitrary initial $\hat{P}_0 = (1, 0.5)$ via each covariate in turn to the final full conditional law. For a binary outcome, this can be done easily using the L'Abbé plots (L'Abbé et al., 1987) shown here, which are simply plots of output against input risks, that is the outcome risk $P(Y = 1)$ after and before passing through a particular flow, respectively. The shapes of the lines illustrate the form of the flow by showing how all possible input risks would be transformed at the relevant estimated size parameter for the flow, and the points illustrate which risks are relevant for the data at hand.

Finally, Figure 7 shows the final L'Abbé plot (representing the treatment effect) for all five models.

### 5.2.3. *Remarks*

From Model 1 we obtain an estimated conditional odds ratio for the treatment effect of 0.560 (95%CI 0.444–0.707). Converted into a (collapsible) covariate-adjusted risk ratio ()), we obtain an estimated risk ratio of 0.585. Estimating a standard error for this standardized estimator would require either an approximation using the delta method, or bootstrapping. Instead, using regression by composition (Model 2), we obtain an estimate of 0.587 for the covariate-adjusted risk ratio, together with its 95% confidence interval 0.473–0.728.

Note that we would not expect these two estimates to be identical, owing to the fact that the models on which they are based make different assumptions (namely, a constant odds-ratio at different baseline risks and a constant risk-ratio at different baseline risks). We see from Figure 7C, however, that these two models lead to very similar predictions for the range baseline risks in the data, which explains the close agreement between the two.

By comparing the published confidence interval and ours from Model 2, there is little precision gained from taking the baseline covariates into account; this is likely due to the large sample size, and thus the balance achieved across randomized groups in the distribution of these covariates.

Other features of the fitted models are also seen in Figures 5, 6 and 7. First, we note the precision of fitted probabilities in the vicinity of fixed points for the flows that have fixed points. Secondly, note that the estimated treatment transformations for Models 3 (risk difference) and 5 (hybrid relative risk) are similar. However, their 95% confidence intervals are very different, reflecting the greater flexibility of the hybrid relative risk model, and therefore its correspondingly lower precision, especially in regions away from the data. The relative imprecision of Model 5 would be less pronounced if the range of estimated input risks into the treatment flows were wider. When this range is narrow (as it is here) it is reasonable to expect many straight lines with positive gradients to be consistent with the data. The apparent precision of the other models thus follows from the additional parametric restrictions (such as fixed points, or a unit slope) made, which in many applications may not be justified.

Finally, we note that—as expected—-some models (including Model 5) lead to flows for the treatment effect that are not closed. However, we note that Model 5 is 'closed on the data' in the

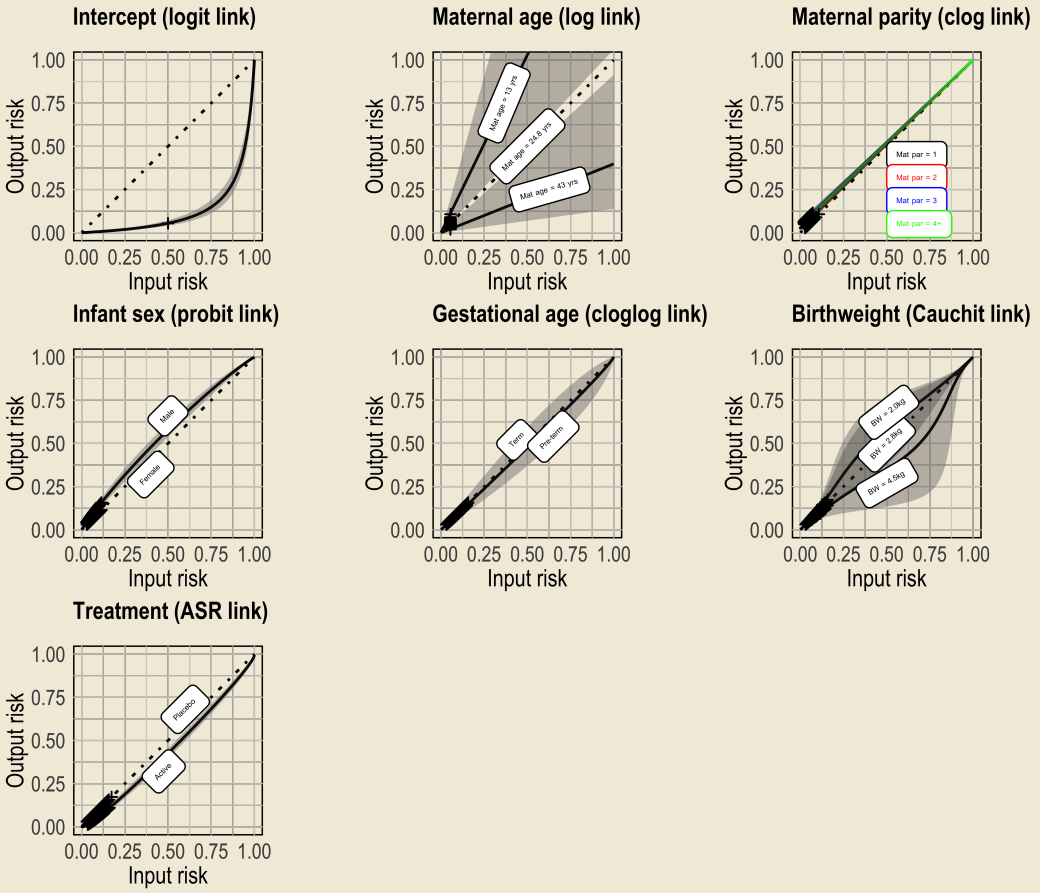


**Fig. 5.** An illustration of the fitted Model 4 as a sequence of L'Abbé plots. For the real-valued covariates (maternal age and birthweight), the 95% confidence intervals for the fitted output risk is plotted at the minimum and maximum values of the covariate (and the pre-standardisation value is given as a label).

sense that the points on the horizontal axis that would be transformed to a negative probability all belong to infants from the placebo group.

## 6. Discussion

The ambition of this paper has been to introduce certain algebraic structures that clarify and expand the scope of regression models, and our understanding thereof. We hope that valuable theory, methodology, and ultimately practical application can be built upon these groundworks. We have largely neglected computational matters, although it is hardly a trivial exercise to compute on survivor functions or other infinite-dimensional representations of distributions. We turn briefly to similarly speculative matters. Some of these merit much more than the passing comment we give them, while others presumably deserve the opposite treatment.

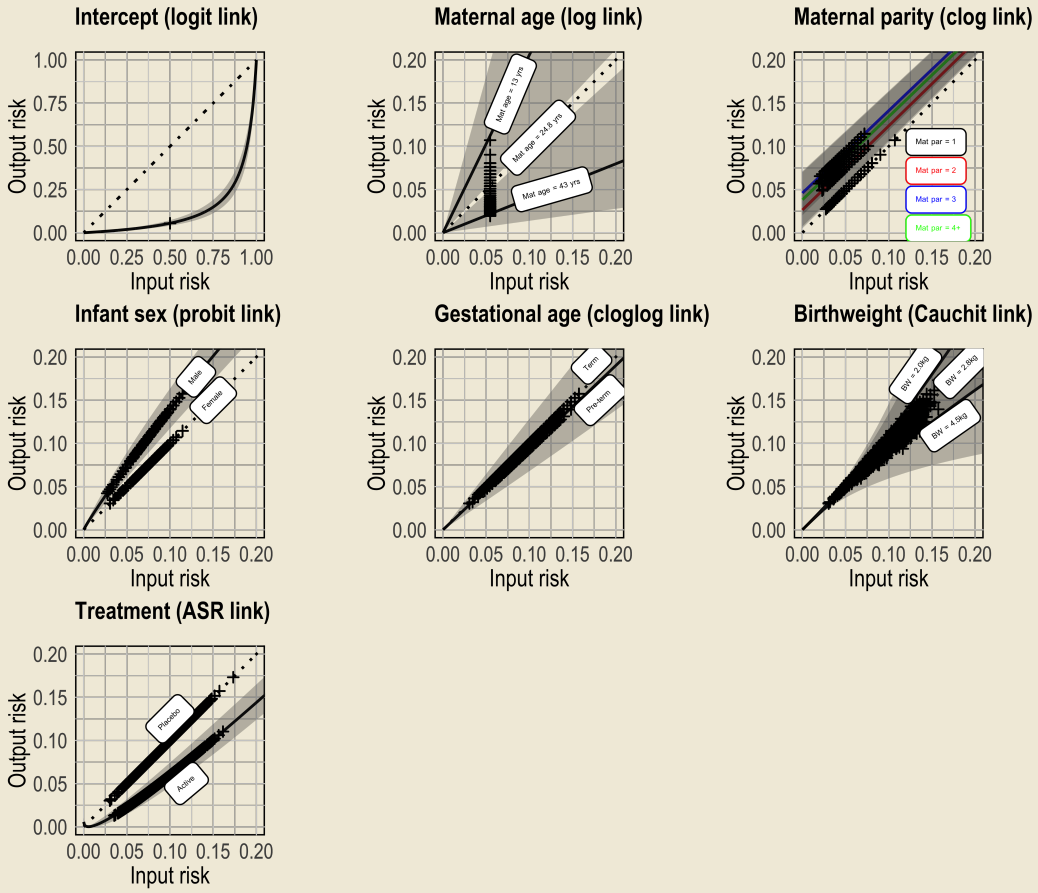


**Fig. 6.** This is the same as Figure 5 except that L'Abbé plots 2–7 are plotted only for input and output risks $\leq 0.2$, since this is where the data lie.

**Table 3.**
The estimated parameters and their standard errors for the five models f

| Covariate | Level | Model 1 | | | Model 2 | | | Model 3 | | | Model 4 | | | Model 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GL link | Est. | SE | GL link | Est. | SE | GL link | Est. | SE | GL link | Est. | SE | GL link | Est. | SE |
| Intercept | | logit | −2.888 | 0.148 | logit | −2.897 | 0.150 | logit | −2.754 | 0.133 | logit | −2.852 | 0.157 | logit | −2.798 | 0.187 |
| Mat. age | | logit | −0.234 | 0.083 | logit | −0.236 | 0.085 | logit | −0.156 | 0.058 | log | −0.219 | 0.102 | logit | −0.180 | 0.096 |
| Parity | 2 | logit | 0.418 | 0.154 | logit | 0.421 | 0.156 | logit | 0.336 | 0.115 | clog | −0.027 | 0.010 | logit | 0.366 | 0.153 |
| | 3 | logit | 0.708 | 0.197 | logit | 0.710 | 0.201 | logit | 0.567 | 0.146 | clog | −0.047 | 0.014 | logit | 0.618 | 0.215 |
| | ≥4 | logit | 0.664 | 0.284 | logit | 0.667 | 0.289 | logit | 0.554 | 0.205 | clog | −0.039 | 0.017 | logit | 0.601 | 0.261 |
| Sex | Male | logit | 0.446 | 0.118 | logit | 0.455 | 0.120 | logit | 0.337 | 0.094 | Probit | 0.198 | 0.053 | logit | 0.372 | 0.144 |
| Gest. age | | logit | −0.068 | 0.197 | logit | −0.065 | 0.200 | logit | −0.106 | 0.161 | cloglog | −0.068 | 0.170 | logit | −0.098 | 0.176 |
| Birthweight | | logit | −0.039 | 0.059 | logit | −0.039 | 0.060 | logit | −0.044 | 0.047 | Cauchit | −0.142 | 0.163 | logit | −0.044 | 0.051 |
| Treatment | Active | logit | −0.580 | 0.119 | log | −0.533 | 0.110 | identity | −0.037 | 0.007 | ASR | −0.075 | 0.015 | log<br>clog | −0.173<br>0.025 | 0.410<br>0.033 |

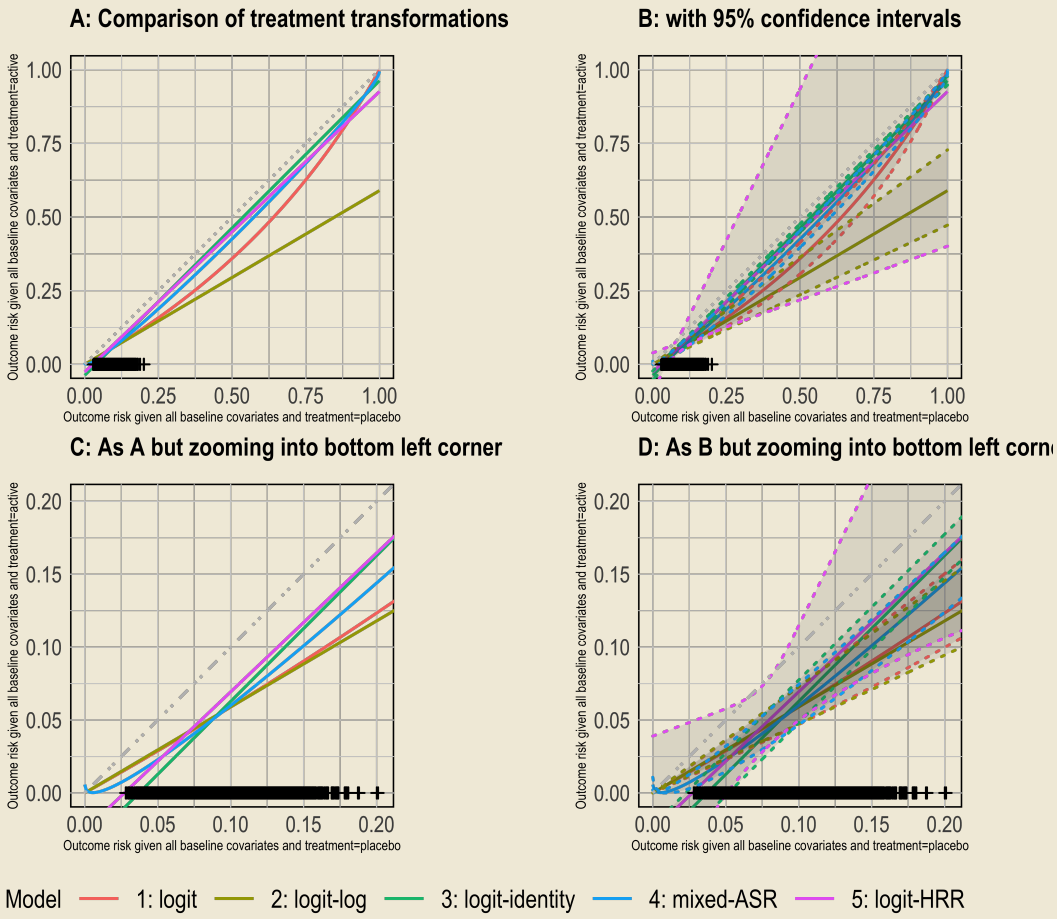


**Fig. 7.** A comparison of L'Abbé plots for the final flow (the treatment transformation), comparing Models 1–5. The input risks (for all models, since these are not exactly the same) are plotted along the horizontal axis.

#### 6.0.1. *Transformations*

Transformations lie at the heart of regressions by composition. The idea of directly estimating transformations is not new, dating back at least to the monograph of Fraser (1968); for an accessible introduction, see Farewell and Prentice (1975). More recently, Hothorn et al. (2018) describe regression parametrizations of transformations applied to cumulative ditribution functions.

One important distinction from previous work is that, in a regression by composition, it is not the *data* that is directly transformed (say by a logarithmic transformation) but instead the *distribution* of the data. We find this distinction to be valuable in practice: we retain the original scale, units (if any) and interpretation of the outcome $Y$, but can nevertheless characterize transformations that are additive, multiplicative, or of a more complicated type. We can also envisage a mixture of transformations, something that is not possible if transformations are done on the data directly.

### *6.1. Other transformation groups*

Regression by composition encompasses many existing models. We do not, however, claim that all regressions can be expressed in this form, nor that it is the most natural form for those that can be so-expressed. However, the broader idea of a compositional construction based on transforming distributions does seem general enough to plausibly include most models that might be designated as regression. We now outline briefly why we believe this to be the case.

The action of a flow $f$ on $\mathcal{M}$ is a special case of the action of a *Lie group*, a continuous group that is a also a differentiable manifold (see, for example, Lee, 2013, pp. 150 sqq.). Vector-indexed flow groups are isomorphic to the simplest kind of differentiable manifold: finite-dimensional vector space. Actions of Lie groups of transformations that are *not* isomorphic to a vector space may also be of interest, and analogous questions also arise concerning the contrast spaces $\mathbb{U}$: it may be that covariate comparisons of interest cannot naturally be embedded in a vector space.

Here we offer two examples of Lie groups $\mathbb{V}$ that are not vector spaces. The first explores the case where the set of transformations being considered is naturally *compact* (which roughly means *closed* and *bounded*). An analogue of the location shift flow for directional data (Fisher, 1953; Fisher, 1995) could be characterized by a rotation of the input probability distribution, and indexed by a complex number $z$ lying on the unit circle $\{z : |z| = 1\} \subset \mathbb{C}$. Because it is compact, the unit circle cannot be given the structure of a vector space, but it is a differentiable manifold that can be equipped with a Lie group structure. This same cyclic structure was apparent in the arcsine-square-root transformation mentioned in Section 4.2.1.

The second example extends the flow idea to noncommutative settings. As we have seen, a dual-flow model combines risk ratio and survival ratio flows, and together they span a set of invertible affine transformations of $\mathcal{M}$. This set $\mathrm{Aff}(\mathcal{M})$ of transformations is a Lie group called the *affine group*. Since affine transformations do not commute in general, this group of transformations cannot be isomorphic to a vector space. However, it does seem more natural to consider $\mathrm{Aff}(\mathcal{M})$ as a single 'flow' rather than two and, because of its differentiable structure, the likelihood derivatives we require may still be defined. Similarly, the accelerated failure time flow (Section 4.1.1) has a natural index space that is not commutative.

Even if the flow is compact or noncommutative, we still need only specify a parametrized linear map from the covariate embedding space into the relevant index group—if an intermediate index space is needed—or directly into a Lie group of transformations. This perspective opens up the beguiling possibility that compositions of the form $f_1 \cdots f_m$ might themselves be elements of

'global' Lie groups if each $f_j$ was an element of a Lie group, a property clearly not shared with flows. Further research in this area therefore seems warranted.

### *6.2. Nonlinear models*

Many of the regressions by composition described in the present paper are nonlinear models, in the specific sense that some of their constituent flows are not affine‡. However, there is a particular sense in which regressions by composition retain the flavour of generalized linear models: they employ a sequence of linear predictors $\eta$ that relate linearly to covariate embeddings $X$ via the estimated linear maps $\theta$.

This linear relationship is convenient for the calculation of derivatives, but could be relaxed to permit parametrized nonlinear mappings between covariate embedding $X \in \mathbb{U}$ and linear predictor $\eta \in \mathbb{V}$. This might be useful, for example, to capture Michaelis–Menten kinematics (Michaelis & Menten, 1913; Johnson & Goody, 2011), for example by setting $\eta(\omega, \theta) = X(\omega)\theta = V(\theta)X(\omega)/\{K(\theta) + X(\omega)\}$, where $V(\theta)$ and $K(\theta)$ are real constants that specify the nonlinear map $\theta\colon \mathbb{U} \to \mathbb{V}$.

Generalized linear models also have semiparametric counterparts that employ robust standard errors. These alternatives allow a modeller to relax the distributional assumption of Poisson-distributed counts, for example. By contrast, the approach to inference for regressions by composition outlined in this paper is fully parametric (with all the attending advantages and disadvantages), so it would be interesting to know whether semiparametric analogues of nonlinear regressions by composition also exist.

### *6.3. Parametric inference?*

Our highly parametric formulation of regression by composition arguably sits somewhat uncomfortably next to efficient nonparametric and semiparametric approaches to inference (van der Laan & Rose, 2011, for example). Putting it bluntly, why should we base inference around a model that is almost surely misspecified, and carries no robust performance guarantees in this case? At first glance, regression by composition seems instead to offer a flexible approach to traditional parametric data modelling: visualize, check distributional assumptions, fit the model, examine diagnostics, and iterate as necessary. Setting aside the potential for insufficient power to detect model inadequacy (Breiman, 2001), the resulting multiple passes through the data should at least give us pause for thought about statistical error control (Leeb & Pötscher, 2005). We offer four tentative suggestions for how to reconcile these ideas.

First, regression by composition enables (and even emphasizes) modelled comparisons of distributions rather than low-dimensional summaries thereof, such as a difference or ratio of means. We contend that distributional comparisons are important in practice in many applications. It is usually not possible to base treatment decisions, say, on knowing only that treated individuals on average fare slightly better than their control counterparts: a modest mean improvement might arise as a mixture of substantial benefit to a small subset of individuals together with moderate harm to the majority. Traditional parametric modelling (with or without regression by composition) has the ambition, and potential, to capture such unanticipated data features. Low-dimensional comparisons do not need drastically differing distributional shapes in order to be of questionable

‡Geert Molenberghs astutely describes the term *generalized linear* (in generalized linear model) as a euphemism (for *nonlinear*).

relevance: the Behrens–Fisher problem (Fisher, 1935) illustrates the challenge of meaningful inference even when only the variance differs.

Second, regressions by composition could be included in ensemble machine learning routines. Certainly regression by composition offers added flexibility to capture explanatory features beyond what is currently possible. Our model

$$\mathbb{P}_m = \mathbb{P}_0 F_1 \cdots F_m$$

already has more than a superficial similarity to a neural network. However, unlike a traditional neural network, in a regression by composition not all covariates need be active in all layers. This has the advantage that the resulting model is likely to be more explainable (see, for example, Giudici & Raffinetti, 2022), and the disadvantage that the user is entirely at liberty to select the 'wrong' covariates for a given layer.

Third, regressions by composition could be used to formulate semiparametric models. Employing a combination of a nonparametric (conceptually, at least) flow and a parametric flow, such semiparametric inference should, of course, account for departures from the incorrect parametric model.

Finally, embracing nonparametric flows throughout a regression by composition seems to us an exciting avenue for future research. For example, it seems appealing to us to be able to produce (possibly smoothed) empirical comparisons of conditional distributions.

### *6.4. Pedagogy*

We have hinted at certain unifying features of regressions by composition. Their constituent transformations have notions of shape and size that are sufficiently intuitive, and at the same time sufficiently abstract, to appeal to us as potentially valuable for teaching purposes. Further, somewhat mysterious relationships between regressions of different outcome types (for instance, Poisson and Cox regression) are slightly easier to fathom when we focus on their corresponding flows. Similarly, apparent paradoxes (such as the collapsibility of accelerated failure flows but not Cox flows despite their coinciding in Weibull regression) can be resolved reasonably neatly.

We also find great value in a notation that emphasizes functions (specifically, transformations of laws) over random variables, analogous to the contrast between Euler and Lagrange's derivative $f'$ and Newton's $\dot{y}$. Function composition draws attention to *local* and *relational* features of models. The temptation with a global decomposition like $Y = \alpha + X\beta + Z\gamma + \epsilon$ is to overinterpret $\epsilon$ as (for example) individual-specific 'measurement error' and, perhaps even more dangerously, the linear predictor $\alpha + X\beta + Z\gamma$ as the 'true $Y$'. In working with laws, regression by composition highlights not the magnitude of the additive discrepancy $\epsilon$ between (say) prediction and observation, but instead the *location* (and hence *surprise*) of the observation in the predicted outcome distribution.

### *6.5. Longitudinal and event-history data*

The present paper offers hints at how basic versions of longitudinal and event-history modelling might be accomplished in a compositional framework. The outcome $Y$ can be a vector of repeated measurements or a stochastic process describing transitions between states. The vector $Y$ might even have a random length to allow for missed clinic visits or early dropout from an observational study (Farewell, 2010). Numbers, timings, and value are all aspects of such a multivariate outcome $Y$ that are amenable to regression by composition modelling. Much better would be a regression

by composition that, like g-methods (Hernan & Robins, 2023, part III), accounts for the possibility of (measured) time-dependent confounding and mediation, and admits a causal interpretation.

**Acknowledgements**

We are grateful to Yacine Trad for insightful questions on an early draft of this paper. We thank Mark Chatfield for helpful discussions about multiplicative comparisons.

**References**

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, *8*(8), 907–925.

Aalen, O. O., Cook, R. J., & Røysland, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis*, *21*(4), 579–593.

Abel, N. H. (1826). Untersuchung der Functionen zweier unabhängig veränderlichen Größen x und y, wie f(x, y), welche die Eigenschaft haben, daß f(z, f (x,y)) eine symmetrische Function von z, x und y ist. *1826*(1), 11–15.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1996, December 20). *Statistical Models Based on Counting Processes*. Springer Science & Business Media.

Aranda-Ordaz, F. J. (1981). On Two Families of Transformations to Additivity for Binary Response Data. *Biometrika*, *68*(2), 357–363.

Bochner, S. (1933). Integration von Funktionen, deren Werte die Elemente eines Vektorraumes sind. *Fundamenta Mathematicae*, *20*(1), 262–176.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society: Series A (General)*, *143*(4), 383–404.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Brown, R. (2018, June 21). *A Modern Introduction to Dynamical Systems*. Oxford University Press.

Calin, O. (2020). Heat-Flow Transformation of a Probability Distribution. *Rev. Roumaine Math. Pures Appl*, *65*(4), 423–438.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Cox, D. R., & Snell, E. J. (1989). *Analysis of Binary Data*. Retrieved February 6, 2023, from https://www.routledge.com/Analysis-of-Binary-Data/Cox-Snell/p/book/9780412306204

Daniel, R., Zhang, J., & Farewell, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, *63*(3), 528–557.

Daniels, H. E. (1954). Saddlepoint Approximations in Statistics. *The Annals of Mathematical Statistics*, *25*(4), 631–650.

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, *21*(11), 1575–1600.

Dieudonné, J. (1960). *Foundations of Modern Analysis*. Academic Press.

Farewell, D. M. (2010). Marginal analyses of longitudinal data with an informative pattern of observations. *Biometrika*, *97*(1), 65.

Farewell, D. M., & Farewell, V. T. (2013). Dirichlet negative multinomial regression for overdispersed correlated count data. *Biostatistics*, *14*(2), 395–404.

Farewell, V. T., & Prentice, R. L. (1975). Interpreting the Structural Model. *Statistische Hefte*, *16*(2), 115–122.

Fine, J. P., Ying, Z., & Wei, L. J. (1998). On the Linear Transformation Model for Censored Data. *Biometrika*, *85*(4), 980–986.

Fisher, N. I. (1995, October 12). *Statistical Analysis of Circular Data*. Cambridge University Press.

Fisher, R. A. (1935). The Fiducial Argument in Statistical Inference. *Annals of Eugenics*, *6*(4), 391–398.

Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *217*(1130), 295–305.

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, *7*(2), 149–154.

Fraser, D. A. S. (1968). *The Structure of Inference*. Wiley.

Giudici, P., & Raffinetti, E. (2022). Explainable AI methods in cyber risk management. *Quality and Reliability Engineering International*, *38*(3), 1318–1326.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Confounding and Collapsibility in Causal Inference. *Statistical Science*, *14*(1), 29–46.

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3), 297–310.

Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, *43*(1), 59–74.

Hernan, M. A., & Robins, J. M. (2023, December 31). *Causal Inference: What If*. CRC Press.

Hernán, M. A. (2010). The Hazards of Hazard Ratios. *Epidemiology*, *21*(1), 13–15.

Hothorn, T., Möst, L., & Bühlmann, P. (2018). Most Likely Transformations. *Scandinavian Journal of Statistics*, *45*(1), 110–134.

Huitfeldt, A., Stensrud, M. J., & Suzuki, E. (2019). On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging Themes in Epidemiology*, *16*(1), 1.

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *106*(4), 620–630.

Johnson, K. A., & Goody, R. S. (2011). The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper. *Biochemistry*, *50*(39), 8264–8269.

Kalbfleisch, J. D., & Prentice, R. L. (2002, September 9). *The Statistical Analysis of Failure Time Data*. Wiley.

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Kreyszig, E. (1978). *Introductory Functional Analysis with Applications*. John Wiley & Sons.

L'Abbé, K. A., Detsky, A. S., & O'rourke, K. (1987). Meta-Analysis in Clinical Research. *Annals of Internal Medicine*, *107*(2), 224–233.

Laird, N. M., & Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, *38*(4), 963–974.

Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*.

Laurent, S. (2012). Some Poisson mixtures distributions with a hyperscale parameter. *Brazilian Journal of Probability and Statistics*, *26*(3), 265–278.

Lee, J. M. (2013). *Introduction to Smooth Manifolds* (Second Edition). Springer.

Leeb, H., & Pötscher, B. M. (2005). MODEL SELECTION AND INFERENCE: FACTS AND FICTION. *Econometric Theory*, *21*(1), 21–59.

Martinussen, T., & Vansteelandt, S. (2013). On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis*, *19*(3), 279–296.

Michaelis, L., & Menten, M. L. (1913). Die kinetik der invertinwirkung Biochem Z 49: 333–369. *Find this article online*.

Nelder, J. A., & Lee, Y. (1991). Generalized linear models for the analysis of taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, *7*(1), 107–120.

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, *7*(4), 308–313.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Nelder, J. A., Lee, Y., Bergman, B., Hynén, A., Huele, A. F., & Engel, J. (1998). Joint Modeling of Mean and Dispersion. *Technometrics*, *40*(2), 168–175.

Nikodym, O. (1930). Sur une généralisation des intégrales de M. J. Radon. *Fundamenta Mathematicae*, *15*, 131–179.

Panigrahi, P., Parida, S., Nanda, N. C., Satpathy, R., Pradhan, L., Chandel, D. S., Baccaglini, L., Mohapatra, A., Mohapatra, S. S., Misra, P. R., Chaudhry, R., Chen, H. H., Johnson, J. A., Morris, J. G., Paneth, N., & Gewolb, I. H. (2017). A randomized synbiotic trial to prevent sepsis among infants in rural India. *Nature*, *548*(7668), 407–412.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, 15–17.

Pearl, J. (2009, September 14). *Causality*. Cambridge University Press.

Pollard, D. (2002). *A user's guide to measure theoretic probability* (Vol. 8). Cambridge University Press.

Richardson, L. F. (1997). The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *210*(459-470), 307–357.

Rücker, G., Schwarzer, G., Carpenter, J., & Olkin, I. (2009). Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*, *28*(5), 721–738.

Scheike, T. H., & Zhang, M.-J. (2002). An Additive-Multiplicative Cox-Aalen Regression Model. *Scandinavian Journal of Statistics*, *29*(1), 75–88.

Schoenberg, I. J. (1946a). Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions: Part a.—on the Problem of Smoothing or Graduation. a First Class of Analytic Approximation Formulae. *Quarterly of Applied Mathematics*, *4*(1), 45–99.

Schoenberg, I. J. (1946b). Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions: Part B—on the Problem of Osculatory Interpolation. a Second Class of Analytic Approximation Formulae. *Quarterly of Applied Mathematics*, *4*(2), 112–141.

Sjölander, A., Dahlqwist, E., & Zetterqvist, J. (2016). A Note on the Noncollapsibility of Rate Differences and Rate Ratios. *Epidemiology*, *27*(3), 356–359.

Thompson, R., & Baker, R. J. (1981). Composite Link Functions in Generalized Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *30*(2), 125–131.

van der Laan, M. J., Hubbard, A., & Jewell, N. P. (2007). Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 463–482.

van der Laan, M. J., & Rose, S. (2011, June 17). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, *11*(14-15), 1871–1879.

Weisberg, S. (2005, April 1). *Applied Linear Regression*. John Wiley & Sons.

Yates, F. (1955). The Use of Transformations and Maximum Likelihood in the Analysis of Quantal Experimients Involving Two Treatments. *Biometrika*, *42*(3-4), 382–403.

Zeng, D., & Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 507–564.