On Estimating Relative Risk

Hansen Zhang

Advisor: Thomas Richardson

August 2025

Abstract

The Relative Risk (RR) is an interpretable parameter in epidemiology and biostatistics, based on both binary input and outcome. It is frequently used in vaccine development to measure the relative efficacy between two treatment groups.

Researchers are often tempted to use generalized linear models (GLMs) to estimate the logarithmic RR as a function of a set of baseline covariates. However, this approach has inherent flaws, as GLMs do not account for variation dependence between the Relative Risk and the Baseline Risk. Richardson et al. have developed an unconstrained and variation-independent nuisance model using the log Odds-Product (OP). They estimate the parameters via maximum likelihood as well as doubly robust semi-parametric approaches.

Building on this work Pozza et al. propose a new nuisance model, but only consider likelihood based estimation. In this work we explore applying doubly robust estimation with the new nuisance model.

Contents

1 Introduction

There are few expository accounts aimed at students which motivate the estimation of relative risk in a practical setting. This paper aims to fill in assumed knowledge and theory behind the techniques utilized in this complex task and explain the story of the estimation of relative risk through academic discourse between Richardson et al. and Pozza et al. Through examples and non-trivial proofs, we reconstruct the history of relative risk models.

2 Model Specification & Theory

The prediction of relative risk (output) with respect to relevant medical characteristics (input) can be modeled using a generalized linear model (GLM). To model it robustly, however, requires further investigation and constraints, such as the intentional design of a nuisance model and another layer of insurance in the form of a doubly robust model.

This section elaborates on why the specification of a nuisance model is required to consistently achieve non-erroneous estimates. In addition, we demonstrate how these models are theoretically specified using the Pozza et al. (9) model as an example.

Complications such as variation dependence, numerical stability issues and model collapsibility are often overlooked by non-specialist researchers who utilize these models, but which we will address in order to improve methods for the estimation of relative risk.

2.1 Defining the problem qualitatively

We begin our investigation by qualitatively describing the problem, or, in other words, understanding the *mathematical territory* of this particular modeling problem.

Remark 1 (A binary treatment-outcome context)**.** The problem is restricted to any scenario where we predict a binary output with respect to a binary input (in addition to confounding variables). This is summarized below with a secondary purpose of introducing some basic notational conventions:

i. The explanatory (independent) variable of interest (A) is binary.

ii. The explained (dependent) variable (Y) is binary.

iii. The confounding variables (V) are not restricted to binary outcomes and can follow any realistic distribution.

iv. Both the support and range take values strictly within the set $\{0, 1\}$ and

v. the expectation $\mathbb{E}[Y|A = a] = p_a \in [0, 1]$

Thus, the modeling problem is of a *binary treatment-outcome* type.

2.2 The significance of relative risk

Now we can move to define the problem epidemiologically, namely to motivate why we care about the parameter of interest–the relative risk (RR).

Suppose that during a clinical trial the investigator wishes to *quantify* the proportional difference between the rate of incidence in the treatment group versus the control group, after already having found that there is a meaningful difference (via a standard statistical test for differences); the relative risk exists to answer this question. Put more specifically, relative risk answers the question:

> How many times more likely is the average treated individual to experience an event than the average non-treated individual as a function of baseline covariates?

2.2.1 Defining inputs

When measuring relative risk under the context of a widespread health initiative or intervention (such as vaccines or medication), it's desirable to stratify our subjects (data) by a set of medical criteria. That is to say, a patient's health condition, biological markers and genetics etc. sometimes have a remarkably important effect on a patient's response to a treatment. But what we aim to measure is the treatment effect on treatment efficacy, which is in turn quantified by the relative risk–not these irrelevant characteristics!

The stratified relative risk reduces noise in the response by clustering based on such shared characteristics. This helps us evaluate whether a certain demographic is an exception, or if a strata needs more data or further investigation in order to improve the odds of success for all populations during an intervention. These medical characteristics are known as *confounding variables* in an epidemiological sense. Mathematically, we will call them *covariates*:

Notation. Let $V \in \mathbb{R}^{n\times p}$ denote a matrix of p covariates belonging to n subjects, and $V = v'$ be a realized stratum of the population.

Notation. Let $A \in \{0, 1\}$ denote a Bernoulli random variable that encodes the status of a given individual as belonging either to the control (A=0) or the treatment (A=1) group.

Remark 2. Although A and V are both covariates, we intentionally segregate them in order to distinguish A as the treatment variable (variable of primary interest) and V as confounding/nuisance variables. Hence, we will let the probability of an event conditioned on *all* covariates be given by:

$$p_a(V) := \mathbb{P}(Y = 1 \mid A = a, V) \tag{1}$$

We call $p_1(V)$ the **exposed risk** and $p_0(V)$ the **baseline risk**.

For any fixed and implicit stratum V (meaning, for clarity, we drop the V), we can also conceptualize the binary outcome as being distributed as:

$$Y|\{A = a\} \sim \text{Ber}(p_a)$$

with mass function:

$$\mathbb{P}(Y \mid A = a) = {p_a}^Y (1 - p_a)^{1-Y}$$

Now that we've laid down notation, we are ready to define relative risk for this paper.

Definition 1 (Relative Risk)**.** *Stratified* relative risk is given by

$$\mathrm{RR}(V) = \frac{P(Y = 1 \mid A = 1, V)}{P(Y = 1 \mid A = 0, V)} = \frac{p_1(V)}{p_0(V)} \tag{2}$$

Example 2.1 (Empirical relative risk)**.** Relative risk is a simple parameter. We illustrate it in an empirical setting through the contingency table below:

	Control (A=0)	**Treatment (A=1)**
Non-Event (Y=0)	A	B
Event (Y=1)	C	D

Then the relative risk in this particular clinical trial is given by:

$$\mathrm{RR}' = \frac{D/(B + D)}{C/(A + C)}$$

where the observed baseline and exposed risk is $p_0' = C/(A + C)$ and $p_1' = D/(B + D)$, respectively. A relative risk < 1 indicates that the average treated individual is *less* likely to experience the event than the non-treated individual. This is the desired conclusion for clinical trials of medications and vaccines. On the other hand, a relative risk > 1 indicates that the treated individual is *more* likely to experience the event than the non-treated individual. This means that, not only does the treatment have no desired effect, it has an adverse effect on the average individual.

Below, we outline some reasons why empirical studies are not always the right choice, and when specifying a model for the relative risk is the right choice.

2.2.2 Motivation for specifying a model for Relative Risk

i. Medical studies have little room for error, leading to high recruitment of subjects and substantial cost and time commitment.

ii. To empirically investigate stratified relative risk, it's tempting to think we can simply marginalize the subjects based on the collected data on their medical characteristics. However, as p grows large, the cardinality of the set of realizations on V grows without bound and is effectively continuous (meaning, the number of combinations of ways we can split subjects is dizzying). Additionally, there is no such rule saying that and individual covariate is discrete to begin. As such, grouping by relative similarity, and hence empirical estimation of relative risk, is not always feasible. We circumvent this problem entirely by modeling relative risk instead.

iii. Another core motivation for modeling relative risk lies in our main goal–prediction–or the convenience of applying historical data. Without any formal experimentation, a researcher can have a very accurate guess of treatment efficacy based on a well-specified model, as well as perform meta-analysis.

iv. avoids cutting corners, i.e using the Odds Ratio (OR) which is easier to model but less interpretable (Richardson et al., 2017).

v. Modeling using stratification leads to variance reduction

2.3 Specifying the relative risk model

Recall that (2) is the quotient of the exposed risk and baseline risk (1) which are themselves the parameters of identically and independently (with respect to A) distributed random variables:

$$Y_i|\{A_i = a\} \overset{\text{i.i.d.}}{\sim} \text{Ber}(p_a) \tag{3}$$

In the following section, we will carefully specify the model that estimates the stratified relative risk using the theory of generalized linear models.

Proposition 1. (3) $\implies p_a(V) = \mathbb{E}[Y|A = a, V]$

Proposition 2. To estimate relative risk conditioned on specific stratum $V = v'$ we let:

$$\log \text{RR}(V) = \log p_1(V) - \log p_0(V) \tag{4}$$

By Proposition 1, we see that the terms which comprise (4) can be modeled by a generalized linear model (GLM) of the form:

$$g\left\{\mathbb{E}[Y \mid A, V]\right\} = A\alpha^T V + \mu^T V$$

where

1. A is the dummy variable encoding group status
2. $(\alpha, \mu) \in \mathbb{R}^{1\times p}$ are coefficient vectors
3. $g(.) := \log(.)$ is the link function of choice

Under the realizations of A discussed in Remark 2, we have the pair of equations:

$$\log p_0(V) = \log \mathbb{E}[Y|A=0,V] = \mu^T V \tag{5}$$
$$\log p_1(V) = \log \mathbb{E}[Y|A=1,V] = \alpha^T V + \mu^T V \tag{6}$$

Taken together, (5) and (6) reconstruct the log relative risk (7) as:

$$\log \mathrm{RR}(V) = \alpha^T V \implies \boxed{\mathrm{RR}(V) = e^{\alpha^T V}} \tag{7}$$

2.3.1 A comment on MLE

In practice, it would be reasonable to first estimate p_0 and p_1 separately as two distinct steps. But deriving Equation 7 as a function of (5) and (6) apriori (*before* estimation) serves two purposes–The first is to separate out the coefficient we care about, α, from the coefficient we don't care about, μ. Secondly, and more importantly, estimating our unknowns (α, μ) jointly produces more efficient estimates due to a larger sample size. If we first estimated p_0 alone, we would be effectively cutting away half the information that MLE uses to learn the coefficient.

However, what *is* true is that the joint surface over (α, μ) is restricted to a subspace in $\mathbb{R}$. This issue will be expounded further in the next section. Then it follows that our nuisance model is characterized by (5) and our parameter model by (7) such that:

> to perform maximum likelihood estimation, we require both models– the relative risk model and the nuisance model.

In effect, (α, μ) become proxy parameters, with μ being the nuisance parameter that is related to p_0.

To summarize, it is tempting to wonder if we could use (7) on its own, but again, we cannot estimate this model on its own via maximum likelihood. We need to estimate μ and α jointly:

$$L(\alpha, \mu;\ \{Y, A, V\}) = \sum_{i=1}^{n} y_i \log(p_a) + (1 - y_i) \log(1 - p_a)$$

Hence, the likelihood function we wish to optimize is a hyper-surface in $\mathbb{R}^3$, which is a function of α and μ. In other words, estimating α cannot escape estimating μ.

2.4 Reconsidering the GLM nuisance model

There are several major issues with the GLM nuisance model that have already been laid out by Richardson et al. (2017), Pozza et al. (2023) and Lumley et al. (2006). We contribute by discussing them in detail for a more general audience.

Remark 3 (Problems of the GLM approach)**.**
Problems I-III arise during the process of decomposing the model estimates (that our nuisance and relative risk model spit out) into it's marginal likelihood:

$$(\log \mathrm{RR}, \log p_0) \to (p_0, p_1)$$

This decomposition helps one decide if the relative risk estimate is reasonable ($\log \mathrm{RR}(V)$ is correctly specified). As such, these problems exist irregardless of whether our sole aim is to estimate relative risk (without also making predictions on the marginal likelihood). However, it can be desirable to the researcher to be able to make predictions on the marginal risk probabilities.

We begin by fixing arbitrary stratum $V = v'$.

I. **(Variation Dependence)**

In short, the condition $(p_0, p_1) \in [0, 1]$ *forces the range of our model to be smaller than* $\mathbb{R}^2$*.*

While it is true that, functionally, p_0 varies dependently on RR (and vice versa), in practice, we construct the relative risk by taking empirical estimates $(\hat{p_0}, \hat{p_1})$ (as we saw in Example 2.1). However, modeling

constraints force us to do the reverse, where we obtain an estimate of the relative risk before inferring the marginal likelihood.

In this sense, the modeling process does not respect that *any* two risk probabilities (p_0, p_1) should be able to produce an estimate of the relative risk. In other words:

> Variation dependence is a natural consequence of the relative risk parameter, which the modeling process does not inherently respect.

As a thought experiment, let's suppose we choose $(\log \mathrm{RR}', \log p_0 = 0.65)$ and want to map it to (p_0, p_1). This is easy with the very simple map given by (5) and (6). As we can see, selection for an estimate on the exposed risk p_1 is restricted to the line $\hat{p_1}(\log \mathrm{RR}) = 0.65 \times e^{\log \mathrm{RR}}$ to ensure the condition $p_1 \in [0, 1]$ is satisfied. But in reality, any choice of RR should give rise to the entire joint likelihood space (unconditional of V). The same problem extends to the coefficient estimates.

II. **(Nonsense Predictions)**

Even if we manage to correctly specify the model through imposing a restriction on the range, the model can still make predictions outside the probability support.

If we let it take liberties, the model will predict outside the probability range on new data. This occurs when the joint range of the covariate distribution is unrestricted:

$$\log(p_1) = u^T V + \alpha^T V = (\mu^T + \alpha^T) V \tag{8}$$

Notice that the left-hand side must be negative for p_1 to be a valid probability. Therefore, the estimate is nonsense if either μ or α is large enough. If V is unbounded, the estimate is also nonsense.

III. **(Convergence Issues & Numerical Instability)** *If the optimum is on the boundary of valid range, then convergence issues and computational instability may arise.* For example, in cases where p_a is very close to 0 or 1, computer algorithms will fail due to numerical precision limitations due to exponential terms. More importantly, algorithms will

explore outside the valid probability range without the aforementioned proper enforcement, which leads to `NaN` values.

In summary, if we do impose a restriction on the range, we get restricted access to our marginal likelihood. If we don't we get predictions outside of $(0, 1)$. If we manage to correctly specify our model, predictions (on new data) can lie outside the proper range.

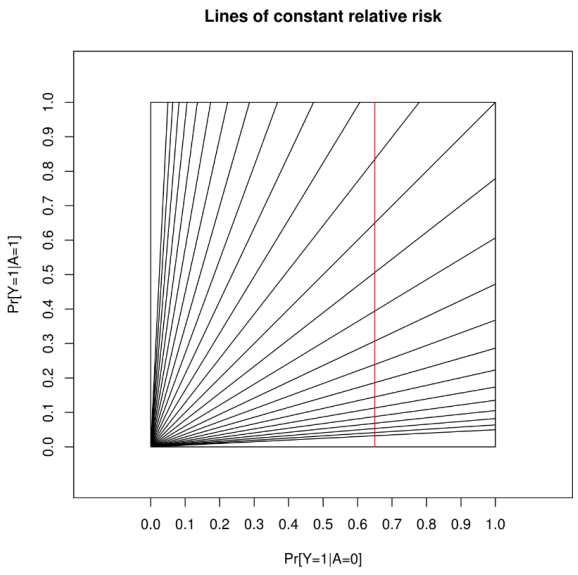

Figure 1: This plot demonstrates Problem (I); Figure 1a in Richardson et al., 2017

We will now take a look at these problems in a practical setting through some examples in R.

Example 2.2 (A tutorial in R)**.** In this simulation (R Core Team, 2024), we first establish our true parameters $\alpha,\ \mu,\ \gamma$, where μ are the coefficients for the nuisance model, α the coefficients for the relative risk model, and γ the coefficients for the propensity score model (which we will define later). All three of these vectors of coefficients depend on our vector of covariates V.

Problem (I)

Under the log-binomial relative risk regression model, we find in our simulation that if we randomly assign a value for the coefficients, we get misspecification of our model a priori. This is because our coefficient parameters are unbounded.

We begin by randomly choosing coefficient values for α and μ:

```
set.seed(1)
n = 400
alpha = c(0.5, 1)
mu = c(1.25, 2)
v0 = rep(1, n)
v1 = rnorm(n, 175, 7) #height in cm
v = cbind(x0,x1)
z = v #assume that z=x
```

We assume a few things: the sample size is $n = 400$, we have one covariate (height) and Z is just the identity transformation on V. Notice that the treatment effect is encoded in $Z^T\alpha$

```
p0 = exp(v %*% mu) #baseline risk
p1 = exp(v %*% mu + z %*% alpha) #treatment risk
```

The true, simulated values of p_0 are outside the range $[0, 1]$. This misspecification immediately disqualifies this model. This misspecification is due to both the randomly assignment of coefficient values, and the height predictor, which is a continuous random variable with no constraint on the support.

In the following fixed version, we proceed with a simulation since restricting our true coefficient values produces reasonable values of (p_0, p_1). We then fit a model to our simulated data:

```
    alpha = c(0.5, -0.01)
    mu = c(1.25, -0.02)
    gamma = c(0.5, -0.007)

    propensity =  exp(v %*% gamma) / (1+exp(v %*% gamma)
       )

    p0 = exp(v %*% mu) #baseline risk
    p1 = exp(v %*% mu+ z %*% alpha) #treatment risk

    a = rbinom(n, 1, propensity)
    pA = numeric(100)
    pA = ifelse(a==1, p1, p0)
    y = rbinom(n, 1, pA)

    data.m = data.frame("event" = y, "intercept" = v0, "
       height" = v1, "treatment" = a)
    bin.fit = logbin(y ~ x1 + a)
    table(round(bin.fit$fitted.values,2))
```

All fitted values are probabilities (between 0 and 1). However, this took tedious consideration of our model parameters, which is not desirable for real world application. Additionally, We will see that this is not always the case if we make estimates with a new set of data, especially if we have outliers values within our data.

Problem (II)

Even if the model is correctly specified, the model can predict values outside the range (0,1) on a new set of data. For example, below we have a point where the height in cm is negative.

```
new_df = data.frame(a = 1, x1 = -100)
x = predict.logbin(bin.fit, new_df, checkminmax = F)
exp(x)
```

Ther resulting prediction for p_1 (exposed risk), is greater than 1.

Problem (III)

Suppose that our data contains 1000 subjects. Now suppose that 999 of them experienced and event ($Y = 1$) while only one does not ($Y = 0$). The covariate V will simply be distributed as the standard normal (so it doesn't really mean anything since it's just noise). For the sake of simplicity, we will fix A=0 because it is enough to show that the GLM for the baseline risk has numerical stability issues.

```

    set.seed(1)
    n <- 1000
    y <- c(0, rep(1, n - 1))              # 999 ones and
       a single zero
    v <- rnorm(n)                         # standard
       normal covariate
    data <- data.frame(y = y, V = V)

    loglik <- function(mu) {
        v <- data$V
        y <- data$y
        p <- 1 - exp(mu * v)
        if (any(p <= 0)) return(-Inf)     # to avoid log
            of non-positive
            sum(y * mu * v + (1 - y) * log(p))
        }

    opt <- optimize(loglik, interval = c(-5, 5), maximum
        = TRUE)

    opt$maximum  # estimated mu
    opt$objective  # on boundary
```

Notice that without the `if` statement, the optimizer would fail due to taking the log of negative values. Even with it included, our optimum μ^* is `Inf`.

Problem (IV) will be addressed in later sections.

3 Pozza nuisance model specification

3.1 The Pozza Nuisance Model

To solve Problems I, II and III all at once, we choose to discard the nuisance model $\log p_0$ and replace it with another choice that allows p_0 (and the marginal likelihood) to be variation independent of a choice of RR. In turn, this allows us to jointly estimate (α, β) via an unconstrained likelihood surface. The problem in question is summarized below:

> We must come up with a function such that every fixed ϕ in the family of curves intersects the entire set which is the family of lines of constant relative risk.

Remark 4. Let V vary implicitly. We illustrate the above point by imagining the cartesian plane of (p_0, p_1), such that p_1 is a function of p_0 with *fixed* θ (or ϕ):

$$p_1{}^{(\theta)} := f(p_0; \theta) \in \mathcal{F}_\theta$$
$$p_1{}^{(\phi)} := f(p_0; \phi) \in \mathcal{F}_\phi$$

Then $\mathcal{F}_\cdot$ is the set of all functions of p_0 and fixed coefficient $(\cdot)$

Put simply, there exists an infinite set of lines of constant relative risk, and an infinite set of curves of constant partial odds product. Mathematically, we wish that:

$$\exists p_0 * \in [0,1] \quad \text{for all} \quad \mathcal{F}_\theta \times \mathcal{F}_\phi = \left\{ \left(p_1{}^{(\theta)}, p_1{}^{(\phi)}\right) \mid p_1{}^{(\theta)} \in \mathcal{F}_\theta,\ p_1{}^{(\phi)} \in \mathcal{F}_\phi \right\}$$

Or that there exists a solution p_0 within the unit square in $\mathbb{R}^2$ for every pairwise function in both families. This implies a closed-form mapping from $(\theta, \phi) \to (p_0, p_1)$, where we've transformed a pair of parameters with range $(0, \infty]^2$ to $[0, 1]^2$.

Many nuisance models can achieve the aforementioned goal, but we will illustrate with the target nuisance model of this paper, the Pozza et al. model.

Definition 2 (Pozza Nuisance Model)**.** The Pozza alternative nuisance model, which we will refer to as the log partial Odds-Product (pOP) is given below:

$$\log \text{pOP}(V) := \log \frac{p_0(V)}{(1-p_0(V))(1-p_1(V))} \tag{9}$$

and

$$\log \text{pOP}(V) = \beta^T V \tag{10}$$

Now that we see the concrete function, we can see that the Pozza model is composed of p_0 and p_1, which allows us to perform the mapping procedure in Remark 4.

To simplify and summarize our system of equations, we will henceforth let $\theta(V) := \log \text{RR}(V)$ and $\phi(V) := \log \text{pOP}(V)$ (Richardson et al., 2017). Then we have the neatened pair of parametric equations written below.

Notation. For convenience of usage in the rest of the paper, we will adopt the abbreviated notation for our system of equations:

$$\theta(V) = \alpha^T V \tag{11}$$
$$\phi(V) = \beta^T V \tag{12}$$

Occasionally, (V) will be dropped to declutter expressions.

Figure 2 suggests that any choice of RR and pOP will result in a pair (p_0, p_1). In other words, the model is variation independent.

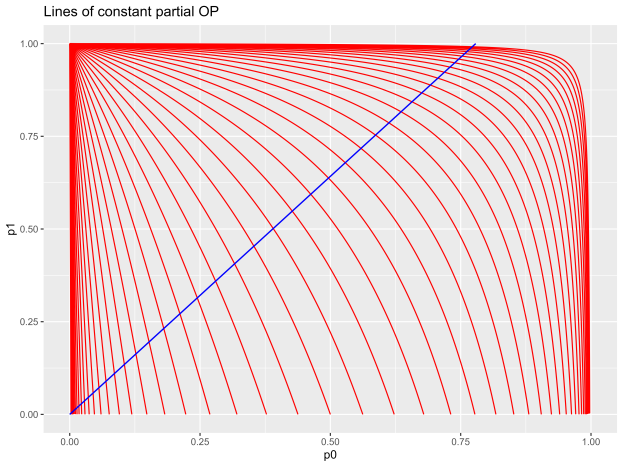

Figure 2: All lines of constant log RR intersect all lines of constant log pOP; generated using the R language (R Core Team, 2024) using `ggplot2`

Theorem 1. Under the log partial Odds-Product, we have the variation independent, closed-form pair of solutions from the mapping:

$$(\theta(V), \phi(V)) \to (p_0(V), p_1(V)) \in [0,1]^2$$

i. $p_1 = p_0 e^{\theta(V)}$

ii. $p_0 = \frac{k+\sqrt{k^2-4e^{\theta(V)+2\phi(V)}}}{2e^{\theta(V)+\phi(V)}}, \quad k = e^{\phi(V)}\left(e^{\theta(V)}+1\right)+1$

Proof.

(Part I) We begin with the trivial solution p_1, which solves equation (7):

$$\begin{aligned} \log(RR) &:= \log\left(\frac{p_1}{p_0}\right) = \theta(V) \\ \Rightarrow \quad e^{\theta(V)} &= \frac{p_1}{p_0}, \quad \boxed{p_1 = p_0 e^{\theta(V)}} \end{aligned}$$

(Part II) Now we substitute p_1 into $\phi(V)$ which gives:

$$\begin{aligned} e^{\phi(V)} &= \frac{p_0(V)}{1-p_0(V)} \cdot \frac{1}{1-p_1(V)} \\ &= \frac{p_0(V)}{[1-p_0(V)][1-p_0(V)e^{\theta(V)}]} \implies \end{aligned}$$

$$\begin{aligned} &e^{\phi(V)}[1-p_0(V)][1-p_0(V)e^{\theta(V)}] = p_0(V) \\ &e^{\phi(V)}\left(1-p_0(V)e^{\theta(V)} - p_0(V) + p_0^2(V)e^{\theta(V)}\right) = p_0(V) \\ &e^{\phi(V)} - e^{\theta(V)+\phi(V)}p_0(V) - e^{\phi(V)}p_0(V) + e^{\theta(V)+\phi(V)}p_0^2(V) - p_0(V) = 0 \end{aligned}$$

which results in the quadratic equation:

$$\left\{e^{\theta(V)+\phi(V)}\right\}p_0^2(V) - \left\{e^{\phi(V)}\left[e^{\theta(V)}+1\right]+1\right\}p_0(V) + e^{\phi(V)} = 0$$

Then using the quadratic formula we get the positive root:

$$\boxed{p_0 = \frac{\left\{e^{\phi(V)}\left[e^{\theta(V)}+1\right]+1\right\} + \sqrt{\left\{e^{\phi(V)}\left[e^{\theta(V)}+1\right]+1\right\}^2 - 4\left\{e^{\theta(V)+2\phi(V)}\right\}}}{2e^{\theta(V)+\phi(V)}}}$$

□

3.2 Alternative nuisance models

We've previously considered the importance of the full (marginal) likelihood and the problems which stem from obtaining it. If one so chooses, they may choose a nuisance model which produces consistent estimates of relative risk, but does not give full likelihood. This leads us into a small tangent in our story, where we explore another variant on the nuisance model which does exactly that. To spell it out, we have three classes of models:

1. full likelihood; variation dependent
2. full likelihood; variation independent
3. no unique likelihood; variation independent

We will explore the third nuisance model class.

Remark 5. If a model cannot provide a unique inverse map (bijection) to (p_0, p_1), it does not give rise to the full likelihood. The consequence is that we cannot make predictions of relative risk based on our covariate A.

Note, we can still estimate relative risk given covariates (we still "learn" that model), but we are not able to say anything about the rate of incidence p_a in a certain stratum V' and simulate individual outcomes under interventions.

One such alternative nuisance model is Tchetgen Tchetgen (2013)'s nuisance parameter $\tilde{\phi}(V) = \log(P(Y=1|A=0,V)/P(Y=1|A=0))$.

Definition 3 (Tchetgen Tchetgen's Parameter)**.** Tchetgen Tchetgen's alternative nuisance parameter is

$$\tilde{\phi}(V) = \log \frac{P(Y=1|A=1,V)}{P(Y=1|A=0,V)} = \log Q(V)$$

Proposition 3. Let $V \in \{0,1\}$ represent binary strata. Then, there exist

$$s_1 = \{P_1(Y=1|A=1,V=v), P_1(Y=1|A=0,V=v)\}$$
$$s_2 = \{P_2(Y=1|A=1,V=v), P_2(Y=1|A=0,V=v)\}$$

where $s_1 \neq s_2$, such that

$$Q_1(V) = Q_2(V), \ \forall V = v$$

See Appendix 7 for a proof of why Tchetgen Tchetgen's nuisance model does not give rise to the full likelihood.

3.3 Developing a new getProbScalar function in brm for Pozza

The bijective mapping for

$$(\theta(V), \phi(V)) \to (p_0, p_1)$$

is theoretically straightforward through solving a system of equations. However, implementation for its computational component is not so straightforward due to numerical instability. To illustrate, we notice that at very large values of $\theta(V) > |12|$, we have that $e^{\theta(V)} \to +\infty$. Numerical precision limitations cannot handle these large values.

To remedy this issue, our goal is to find the limit of p_0 as one of the parameter $(\theta(V), \phi(V))$ diverges to $\{-\infty, +\infty\}$ while the other remains fixed.

Proposition 4. Recall that (7) $\implies p_1(V) = p_0(V)e^{\theta(V)}$

(Case I) "West Edge" $(\theta, \phi) \to (+\infty, -\infty)$

Theorem 2. For the parametric pair $(\theta, \phi) \in \mathbb{R}^2$, if

$$\theta \to \pm\infty, \quad \phi \to \mp\infty$$

Then the limit of the mapping (p_0, p_1) is given by:

$$\begin{aligned} {p_0}^* &= \frac{1+e^y}{2e^y} = \frac{1+e^{\theta+\phi}}{2e^{\theta+\phi}} \\ {p_1}^* &= {p_0}^* \times e^\theta \end{aligned}$$

Proof. Fix $y = \theta + \phi$ for $(\theta, \phi) \in R^2$ such that their sum is constant. In other words, we have the condition $\theta = m\phi$ for some arbitrary $m \in \mathbb{R}$. Then notice that $y = \theta + \phi \implies$

$$\theta = y - \phi \tag{13}$$

We substitute (13) into the mapping of p_0 such that:

$$\begin{aligned} p_0(y) &= \frac{[1+e^\phi(1+e^\theta)] - \sqrt{[1+e^\phi(1+e^\theta)]^2 - 4e^{\theta+2\phi}}}{2e^{\theta+\phi}} \\ &= \frac{[1+e^\phi(1+e^{y-\phi})] - \sqrt{[1+e^\phi(1+e^{1-\phi})]^2 - 4e^{y+\phi}}}{2e^y} \\ &= \frac{[1+e^\phi+e^y] - \sqrt{[1+e^\phi+e^y]^2 - 4e^{y+\phi}}}{2e^y} \end{aligned}$$

Then taking the limit with respect to ϕ results in the expression:

$$\lim_{\phi\to\infty} p_0(y) = \frac{1+e^y}{2e^y} = \frac{1+e^{\theta+\phi}}{2e^{\theta+\phi}}$$

□

(Case II) "North Edge" $(\theta > 0,\ \phi \to +\infty)$

Observe that:

$$p_1 \to 1 \implies$$
$$1 = p_0 \times e^{\theta}$$

such that

$$\boxed{(p_0 = e^{-\theta},\ p_1 = 1)}$$

(Case III) "East Edge" $(\theta < 0,\ \phi \to +\infty)$

$$p_0 \to 1 \implies$$
$$p_1 = p_0 \times e^{\theta} = e^{\theta}$$

such that

$$\boxed{(p_0 = 1,\ p_0 = e^{\theta})}$$

The three cases above address all the limiting behavior of the mapping to (p_0, p_1) such that the function will never return `NaN` for either value.

4 Double-robustness

To recap, specifying an alternative nuisance model that solves the variation dependence problem is simple. All we need to do is come up with a function that intersects the entire family of lines of constant relative risk.

Let's now suppose that we misspecify the functional form of the nuisance model, $\phi(V)$, such that our μ estimate is not consistent. Then because maximum likelihood estimation optimizes over a hypersurface in $\mathbb{R}^3$, the inconsistent estimate of β "contaminates" the estimate of α at the joint optimum of the curve (α^*, β^*). This means that each component of the joint optimum is not necessarily equal to the the marginal optimums.

This describes Problem (IV) in a nutshell. The solution is to take a doubly-robust approach. This requires that we discard our potentially inconsistent estimate of α, while keeping our potentially inconsistent estimate of β. Remember, the above paragraph indicates that (α, β) come bundled, where if ϕ is misspecified, both estimates are inconsistent and if not, both estimates are consistent.

We repeat the nuisance model procedure once more, but this time, the key is to use a parameter to build a model that is difficult to misspecify. A natural choice is the propensity score:

Definition 4 (Propensity Score)**.** The propensity score, $e(v) := P(A = 1|V)$, is the proportion of assignment to the treatment group (A=1), conditioned on the stratum V. It is modeled by the logit GLM:

$$logit(e(v)) = \log\left(\frac{e(v)}{1 - e(v)}\right) = \gamma^T V$$

From Definition (4), it is easy to see why the propensity score is difficult to misspecify–it is nothing more than the probability of assignment to to the treatment group. The design of the clinical trial itself naturally gives rise to the propensity score (Richardson et al., 2017).

5 Applying the doubly-robust framework to the Pozza nuisance model

Pozza et al. (2023) specify an alternative nuisance model, which we will call the partial Odds-Product (pOP), but do not attempt to adapt the doubly-robust component of the original Richardson et al. log Odds-Product (OP) model (you can refer to the paper to see the specification), which is a key feature of the latter model. We would now like to implement this doubly robust component.

5.1 Simulation

Remark 6 (Simulation parameters)**.** We will be running the simulation with $n = 2000$ samples and $m = 200$ simulations.

In this section we test the estimation accuracy of the coefficients (α, β) (11) (12) under different practical scenarios which may arise but are less than ideal.

We use bias and standard errors of the proxy parameter (the coefficient α) to test estimation consistency.

The four scenarios we will be testing will include:

- Pozza Model correct (not doubly robust): **MLE**
- Pozza Model incorrect (not doubly robust): **MLE Wrong** ϕ
- Both the nuisance and propensity model are correctly specified: **DR BTH**
- Nuisance model is wrong, propensity model is correct: **DR PSC**

These four will be sufficient to compare our results to the log OP model.

Remark 7. Using the numerically stable function defined in the previous section, and simulation skeleton code from `brm` package, we conduct a test of estimation consistency between the Pozza model and Richardson model.

The results are below, where the results from the Richardson et al., 2017 model are drawn directly from their own simulation.

Table 1: Bias (Standard Error) for Alpha Parameters

Model		**MLE**	**MLE Wrong ϕ**	**DR BTH**	**DR PSC**
Richardson	α_0	0.008 (0.004)	-0.403 (0.004)	0.008 (0.004)	0.007 (0.004)
	α_1	-0.021 (0.005)	0.021 (0.004)	-0.023 (0.005)	-0.022 (0.005)
Pozza	α_0	0.001 (0.006)	-0.448 (0.004)	0.026 (0.007)	0.023 (0.007)
	α_1	-0.024 (0.005)	0.546 (0.003)	-0.097 (0.013)	-0.093 (0.013)

Both the bias and standard error of the Pozza model are similar to that of the Richardson model.

6 Discussion

This paper's main aim is to provide a gentle introduction to estimating relative risk with respect to baseline covariates. Future directions may include specification of a novel alternative nuisance model, and testing the Pozza model on real data.

References

Lumley, T., Kronmal, R., & Ma, S. (2006). Relative risk regression in medical research: Models, contrasts, estimators, and algorithms.

Pozza, F., Pagui, E. C. K., & Salvan, A. (2023). Improved and computationally stable estimation of relative risk regression with one binary exposure [PMID: 37032617]. *Statistical Methods in Medical Research*, *32*(6), 1234–1246. https://doi.org/10.1177/09622802231167436

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Richardson, T. S., Robins, J. M., & Wang, L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, *112*(519), 1121–1130. https://doi.org/10.1080/01621459.2016.1192546

Tchetgen Tchetgen, E. (2013). Estimation of risk ratios in cohort studies with a common outcome: A simple and efficient two-stage approach. *International Journal of Biostatistics*, *9*(2), 251–264. https://doi.org/10.1515/ijb-2013-0007

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

7 Appendix

A. Proof on Tchetgen Tchetgen's nuisance model

Proof. Let our 2-dimensional space be represented by:

$$[0,1] \times [0,1] = \{(x,y) \in R : 0 \leq x, y \leq 1\}$$

where:

$$x = P(Y = 1|A = 0, V') \text{ is the baseline risk}$$
$$y = P(Y = 1|A = 1, V') \text{ is the exposed risk}$$

given the same n-length vector of co-variates V' Assume V is bivariate,

$$V \to f(V) \in \{0,1\}$$

Then conditioning on the two unique strata V_1 and V_2 we have that x and y are always linearly related:

$$RR(V_i) = \frac{P(Y = 1|A = 1, V_i)}{P(Y = 1|A = 0, V_i)}$$
$$P(Y = 1|A = 1, V_i) = RR(V_i)P(Y = 1|A = 0, V_i) \implies$$
$$y = RR(V_i)x$$

Let us now have two scenarios, s_1 and s_2 such that:

$$s_1(V_i) = \{P_1(Y = 1|A = 1, V = v), P_1(Y = 1|A = 0, V = v)\}$$
$$s_2(V_i) = \{P_2(Y = 1|A = 1, V = v), P_2(Y = 1|A = 0, V = v)\}$$

such that $s_1 \neq s_2$ and $V = v'$ specific.

Since x and y are linearly related, any scaling of x will be the same scaling of y. For simplicity sake, assume $s_2 = \alpha s_1$, where $\alpha \in R$ is some constant. Then $s_1 \neq s_2$. The figure below represents the problem thus far geometrically.

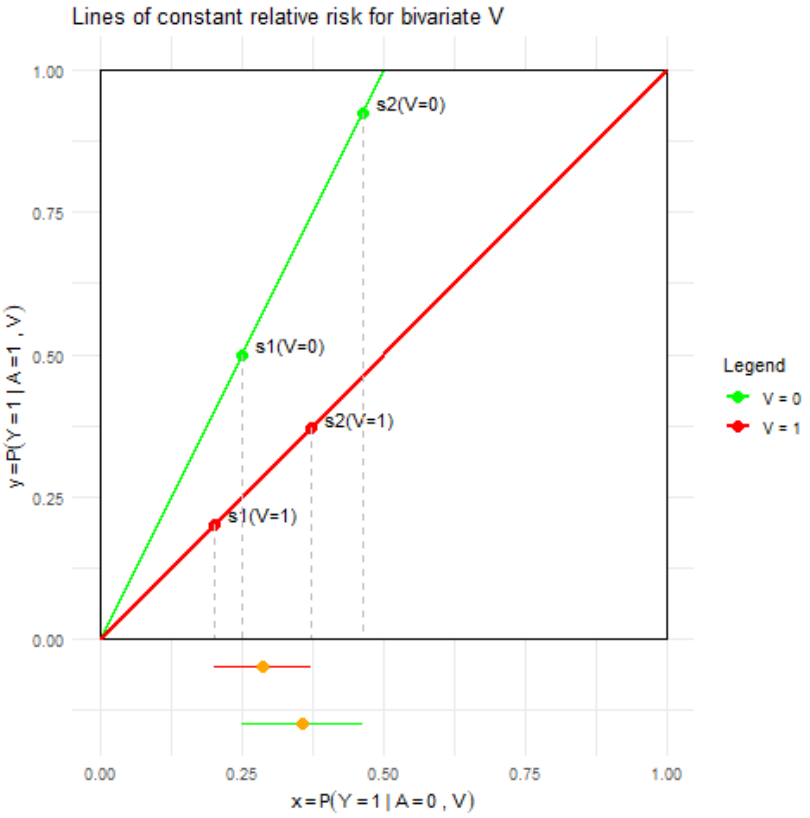

Figure 3: Lines of constant relative risk; If we let V_i be non-specific, we see that

Recall that:

$$\begin{aligned} P(Y=1|A=0) &= P(Y=1, V=0|A=0) + P(Y=1, V=1|A=0) \\ &= P(Y|A=0, V=0)P(V=0|A=0) + \\ &\quad P(Y|A=0, V=1)P(V=1|A=0) \end{aligned}$$

Then we have that

$$
\begin{aligned}
Q_2(V) &= \frac{P_2(Y=1|A=0,V)}{P_2(Y=1|A=0)} \\
&= \frac{P_2(Y=1|A=0,V)}{P_2(Y|A=0,V=0)P(V=0|A=0)+P_2(Y|A=0,V=1)P(V=1|A=0)} \\
&= \frac{\alpha P_1(Y=1|A=0,V)}{\alpha P_1(Y|A=0,V=0)P(V=0|A=0)+\alpha P_1(Y|A=0,V=1)P(V=1|A=0)} \\
&= \frac{P_1(Y=1|A=0,V)}{P_1(Y=1|A=0)} \\
&= \frac{P_1(Y=1|A=0,V)}{P_1(Y=1|A=0)} \\
&= Q_1(V)
\end{aligned}
$$

And hence our proof is complete.

□