The Karlson-Holm-Breen (KHB) Method: Why Logistic Mediation Results Might Be Misleading?

Ibrahim Hasan and S M Abdullah

A simple story with a built-in complicated problem

Assume being a health researcher, you are studying health inequality. Particularly, you are interested to know the non-communicable disease (NCD) outcome status due to educational attainment exposure. Your preliminary logistic regression results showed that people with low educational attainment are more likely to be hypertensive. Since the literature establishes that NCD outcomes depend causally on physical activity,(Lee et. al., 2012) you included physical activity in the model to ensure the analysis is realistic and intuitive. In new set of results, education coefficient, suppose dropped by 30% and accordingly, you concluded "physical inactivity explains about one-third of the education gap in hypertension". Your analysis is methodologically careful and rigorous; nonetheless, there is a great possibility that the "one-third" figure is misleading. This is not due to data quality or flawed study design, but because logistic regression acts differently than many of us assume. Unless you correct for that difference, your mediation results are mixing real effects with a statistical illusion. This is where the Karlson-Holm-Breen (KHB) method comes into play.(Karlson, Holm and Breen, 2012)

Before getting into technical aspects, let's explore why mediation analysis matters in the first place. In public health and social science,

researchers rarely just want to know whether X affects Y. They want to know "how". For instance, in the scenario mentioned above they often frame questions as "does education reduce hypertension partly because it increases physical activity?" This is what mediation analysis tries to answer. It separates: direct effect of an exposure (X affects Y), and indirect effect operating through a mediator (X affects Y through a mediator M). In linear regression, such decomposition is simple, involving just coefficient comparisons to determine the mediator and the extent of mediation. In contrast, logistic regression presents more complexity.

What most researchers do and why it breaks?

The standard approach for mediation analysis has the following steps:

Step 1:
Estimate Model 1: `logit(Y) ~ X`

Step 2:
Estimate Model 2: `logit(Y) ~ X + Mediator`

Step 3:
Estimate proportion mediated:
$\text{Mediation Proportion (PM)} = (\beta_1 - \beta_2) / \beta_1$

In linear regression, it works perfectly. In practice, researchers augment the model with a variable, examine the coefficient adjustment, and attribute the change to the added variable. The coefficient of X stays meaningful across models. However, in logistic regression, the coefficient of X changes for two reasons; real mediation and a built-in scaling issue called non-collapsibility. Ignoring the latter reason leads to biased mediation estimation.

The non-collapsibility problem

A statistical measure is collapsible if it remains stable when conditioned on an additional variable, provided that variable is not a confounder. Augmenting a linear regression model with a variable unrelated to X is expected to leave the coefficient of X unchanged. In contrast, in logistic regression, the coefficient of X may change even if the added variable is unrelated to X. This change does not indicate a true relationship alteration but reflects a mathematical property of odds ratios: they are non-collapsible. Logistic regression coefficients are not directly comparable across nested models, even with the same data. Therefore, when researchers observe coefficient changes after adding a mediator, part of this may simply be a scaling artifact.

The reason: a fixed hidden variance

Logistic regression assumes an unobserved continuous variable underlying the binary outcome. The variance of this latent scale is fixed at $\pi^2/3 \approx 3.29$ (the variance of the standard logistic distribution), and this does not change regardless of what variables the researcher adds to the model.

In linear regression, residual variance is estimated freely from the data. Adding a new variable, leads to shrink in the residual variance, while the other coefficients remain stable. In logistic regression, the total variance is fixed at $\pi^2/3$. For adding a mediator that explains some of that variance, the residual variance must shrink. But because the total variance is fixed, the model rescales the remaining coefficients upward. This is why naive mediation in logistic regression often overstates indirect effects and can even create artificial suppression effects. Winship and Mare noted this in 1984.(Winship and Mare, 1984)

The KHB solution: residualise the mediator

Karlson, Holm, and Breen proposed a solution of this problem in 2012.(Karlson, Holm and Breen, 2012) Instead of adding the mediator (M) directly to the logistic model, first residualise it (regress M on the exposure X) and retain the residuals $\tilde{M}$. This residualised variable is, by construction, unrelated to X: all variance in M attributable to X has been partialled out.

Adding $\tilde{M}$ to the reduced model does not change the coefficient on X, because $\tilde{M}$ is uncorrelated with X, its inclusion does not reduce X's explained variance and therefore does not trigger rescaling. The reduced model

(with M̃) and the full model (with M) are reparameterisations of the same model: they fit identically. But they now share the same residual variance and the same error distribution, making coefficient comparison valid.

R and STATA code

The core R code would look like this:

 # Step 1:
Residualise the mediator on the exposure

```
M_tilde <- residuals(lm(M ~ X + covariates))
```

 # Step 2:
Full model — X + M → Direct Effect (NDE)

```
mod_full <- glm(Y ~ X + M + covariates, family = binomial())
```

 # Step 3:
Reduced model — X + M̃ → Total Effect (scale fixed)

```
mod_red  <- glm(Y ~ X + M_tilde + covariates, family = binomial())
```

 # KHB Decomposition

```
TE  <- coef(mod_red)["X"]    # Total Effect

NDE <- coef(mod_full)["X"]   # Natural Direct Effect

NIE <- TE - NDE              # Natural Indirect Effect (via M)

PM  <- NIE / TE              # Proportion Mediated
```

Fortunately, the KHB method is comparatively simpler to implement in both Stata and R, thereby enhancing its accessibility to most applied researchers.

Karlson, Holm, and Breen have provided the 'khb' command for Stata, available from the SSC archive. Installation and basic use are as follows:

```
ssc install khb

khb logit y x || m, disentangle
```

The disentangle option provides a full decomposition, including the contribution of each mediator when multiple mediators are specified.

In R, the KHB package provides equivalent functionality:

```
install.packages("KHB")

khb(model.full, model.reduced, mediator = "m")
```

Both implementations report total, direct, and indirect effects on the log-odds scale, along with standard errors and confidence intervals derived via the delta method or bootstrapping. Because both models are on an identical scale, the additive decomposition (Total Effect = Direct Effect + Indirect Effect) holds exactly on the log-odds scale, the principal advantage of the KHB approach over conventional methods.

Reporting the Decomposition

The KHB paper (Karlson, Holm and Breen, 2012) presents three equivalent forms of the decomposition, each suited to different reporting conventions:

Method	What it shows
Difference	Raw KHB coefficient change; analogous to standard logit coefficient difference
Ratio	Proportion of total effect mediated; scale-free
Percentage change	Mediated fraction expressed as Mediation Proportion (PM); easiest to communicate to non-specialists

Assumptions and limitations

Like all methods for causal mediation analysis, KHB rests on identifying assumptions that researchers should be explicit about.

- No unmeasured confounding of X→Y: the exposure-outcome relationship should not be confounded by unmeasured variables.
- No unmeasured confounding of M→ Y: the mediator-outcome relationship is similarly unconfounded. This is often the more demanding requirement in practice.
- No X–M interaction: the standard KHB decomposition assumes no interaction between the exposure and the mediator in their effect on the outcome.
- Correct model specification: as with all regression-based methods, misspecification can bias results.

It is also important to note that the KHB decomposition produces estimates on the log-odds scale, which may limit direct substantive interpretation. Researchers who need results on the probability scale (e.g. risk differences) should consider the 'counterfactual causal mediation framework',(VanderWeele, 2015) which produces natural direct and indirect effects with bootstrap confidence intervals on the probability scale. In practice, both approaches can be used simultaneously: KHB as the primary decomposition within the logistic framework,(Karlson, Holm and Breen, 2012) and VanderWeele as a sensitivity check on the probability scale. (VanderWeele, 2015)

The Takeaway

For nearly three decades following Winship and Mare's identification of the rescaling problem,(Winship and Mare, 1984) mediation estimates from logistic regression models were routinely reported without correction for a statistical artefact that could partially, and in some cases substantially, distort conclusions.

The KHB method does not require abandoning logistic regression. It requires residualising the mediator before adding it to the model, which costs nothing in terms of computation but fixes everything in terms of validity. The method is not a substitute for thoughtful causal reasoning; the identifying assumptions remain demanding, and cross-sectional designs cannot confirm causal direction. But it ensures that the arithmetic of effect decomposition is at least internally consistent, and that the proportion mediated reported in published work reflects genuine mediation rather than a scaling artefact.

References

1. Lee IM, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT; Lancet Physical Activity Series Working Group. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet. 2012 Jul 21;380(9838):219-29. doi: 10.1016/S0140-6736(12)61031-9
2. Karlson, Kristian Bernt, Anders Holm, and Richard Breen. 2012. "Comparing regression coefficients between same-sample nested models using logit and probit: A new method." *Sociological Methodology* 286-313.
3. Winship, Christopher, and Robert D. Mare. 1984. "Regression models with ordinal variables." *American Sociological Review* (American Sociological Review) 512-525.
4. VanderWeele, Tyler J. 2015. *Explanation in causal inference: Methods for mediation and interaction.* New York: Oxford University Press.

Ibrahim Hasan
Research Assistant, ARK Foundation

S M Abdullah
Associate Professor, Department of Economics
University of Dhaka