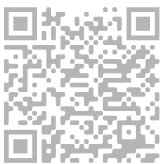
Direct and Indirect Discrimination in Generalized Linear Models

Arthur Charpentier
Université du Québec à Montréal

with **Bertille Tierny** and **François Hu**



7th European Actuarial Journal Conference, Istanbul

What are we trying to decompose?

Let S be a sensitive attribute, $\boldsymbol{X}$ non-sensitive covariates, and $f(\boldsymbol{X}, S)$ a fitted prediction.

Diagnostic question

When prediction distributions differ across groups,

strong demographic (un)parity

$$\mathcal{L}\{f(\boldsymbol{X}, S) \mid S = s\} \neq \mathcal{L}\{f(\boldsymbol{X}, S) \mid S = s'\}$$

where does the disparity come from?

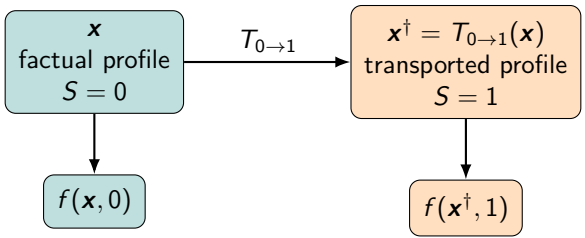
explicit sensitive effect + covariate/proxy channel + distributional structure.

Fairness background:
[Dwork et al., 2012, Hardt et al., 2016, Kusner et al., 2017, Barocas et al., 2017, Charpentier, 2024].

Counterfactual comparison by transport

A transport map compares individuals *mutatis mutandis*:

$$T_{0\to1\#}\,\mathcal{L}(\boldsymbol{X} \mid S=0) = \mathcal{L}(\boldsymbol{X} \mid S=1), \qquad \boldsymbol{x}^\dagger = T_{0\to1}(\boldsymbol{x}).$$



$$\underbrace{f(\boldsymbol{x}^\dagger,0) - f(\boldsymbol{x},0)}_{\text{covariate / indirect}} \qquad \underbrace{f(\boldsymbol{x}^\dagger,1) - f(\boldsymbol{x}^\dagger,0)}_{\text{residual / direct}}.$$

[Villani, 2009, Santambrogio, 2015, Gordaliza et al., 2019, Black et al., 2020, Charpentier et al., 2023, De Lara et al., 2024, Fernandes Machado et al., 2026].

From Oaxaca–Blinder to prediction decompositions

Classical Kitagawa–Oaxaca–Blinder decompositions split a mean gap into characteristics and coefficients:

$$\bar{\boldsymbol{x}}_1^\top \boldsymbol{\beta}_1 - \bar{\boldsymbol{x}}_0^\top \boldsymbol{\beta}_0 = \underbrace{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_0)^\top \boldsymbol{\beta}_0}_{\text{characteristics}} + \underbrace{\bar{\boldsymbol{x}}_1^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)}_{\text{coefficients}}.$$

corresponding to $\mathbb{E}[f(\boldsymbol{X}, S) \mid S = 1] - \mathbb{E}[f(\boldsymbol{X}, S) \mid S = 0]$.

Our use of the idea

We do *not* compare two group-specific regressions. We decompose the disparity generated by a *single fitted predictor*.

$$f_{\text{lin}}(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle + \gamma_s + \beta_0.$$

[Kitagawa, 1955, Oaxaca, 1973, Blinder, 1973, Jann, 2008, Fortin et al., 2011, Fairlie, 2005].

From mean gaps to Wasserstein disparity

Mean gaps miss dispersion and shape. Let

$$\nu_{f|S=s} = \mathcal{L}\{f(\boldsymbol{X}, S) \mid S = s\}.$$

A Wasserstein barycentric measure of distributional disparity is

$$W_2^2(\nu_0, \nu_1) = \int_0^1 (F_0^{-1}(u) - F_1^{-1}(u))^2 du$$

$$U_{W_2}(f) := \min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{M} \pi_s \, W_2^2\big(\nu_{f|S=s}, \nu\big)$$

Linear-Gaussian simplification

If

$$f_{\text{lin}}(\boldsymbol{X}, S) \mid S = s \sim \mathcal{N}(m_s, b_s^2),$$

then

$$U_{W_2}(f_{\text{lin}}) = U_2(f_{\text{lin}}) = \text{Var}_S(m_S) + \text{Var}_S(b_S).$$

[Agueh and Carlier, 2011, Dowson and Landau, 1982, Gelbrich, 1990, Takatsu, 2011, Fukuchi and Sakuma, 2023].

Linear benchmark: exact four-way decomposition

Consider one fitted predictor

$$f_{\text{lin}}(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle + \gamma_s + \beta_0$$

For each group, suppose $\boldsymbol{\mu}_s = \mathbb{E}[\boldsymbol{X} \mid S = s]$, $\boldsymbol{\Sigma}_s = \text{Cov}(\boldsymbol{X} \mid S = s)$ and denote

$$m_s = \mathbb{E}[f_{\text{lin}}(\boldsymbol{X}, S) \mid S = s], b_s = \sqrt{\text{Var}(f_{\text{lin}}(\boldsymbol{X}, S) \mid S = s)}.$$

Under the Gaussian benchmark,

$$U_{W_2}(f_{\text{lin}}) = \text{Var}_S(m_S) + \text{Var}_S(b_S).$$

Since $m_S = \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle + \gamma_S + \beta_0$, we obtain

$$U_2(f_{\text{lin}}) = \underbrace{\text{Var}_S(\gamma_S)}_{\text{direct}} + \underbrace{\text{Var}_S(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{\text{indirect}} + \underbrace{2\,\text{Cov}_S(\gamma_S, \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{\text{interaction}} + \underbrace{\text{Var}_S(b_S)}_{\text{structural}}.$$

[Tierny et al., 2026a, Tierny et al., 2026b].

How to read the four terms

$$U_2(f_{\text{lin}}) = \underbrace{\text{Var}_S(\gamma_S)}_{\text{direct}} + \underbrace{\text{Var}_S(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{\text{indirect}} + \underbrace{2\,\text{Cov}_S(\gamma_S, \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{\text{interaction}} + \underbrace{\text{Var}_S(b_S)}_{\text{structural}}.$$

- **Direct:** explicit fitted contribution of the sensitive attribute.
- **Indirect:** differences in average non-sensitive covariates.
- **Interaction:** direct and indirect channels reinforce or offset.
- **Structural:** differences in within-group predictive dispersion.

Caution

A negative interaction is only a statistical offset in the fitted score, not a normative correction.

Why generalized linear models?

Actuarial targets are rarely Gaussian on their natural scale.

Actuarial quantity	**Typical GLM**
Occurrence / hospitalization	Logistic regression
Frequency / number of visits	Poisson regression
Utilization / pure premium / cost	Tweedie regression

GLMs combine a latent additive score $\eta = \langle \boldsymbol{X}, \boldsymbol{\beta} \rangle + \gamma_S + \beta_0$

with a nonlinear inverse link $f(\boldsymbol{X}, S) = h(\eta)$.

Key issue

A disparity that is additive on the latent scale may be compressed, amplified, or reshaped on the prediction scale.

[McCullagh and Nelder, 1989, De Jong and Heller, 2008, Ohlsson and Johansson, 2010, Frees, 2009, Dunn and Smyth, 2018, Denuit et al., 2019].

The GLM diagnostic: three quantities

For

$$f(\boldsymbol{X}, S) = h(\eta),$$

define the empirical output-scale criterion

$$U_2(f) = \mathrm{Var}_S(\mu_f(S)) + \mathrm{Var}_S(\sigma_f(S)),$$

where

$$\mu_f(s) = \mathbb{E}[f(\boldsymbol{X}, S) \mid S = s], \qquad \sigma_f(s) = \sqrt{\mathrm{Var}(f(\boldsymbol{X}, S) \mid S = s)}.$$

The first-order proxy transports latent moments:

$$\widetilde{\mu}_f(s) = h(m_s), \qquad \widetilde{\sigma}_f(s) = h'(m_s) b_s,$$

and

$$\widetilde{U}_2(f) = \mathrm{Var}_S(\widetilde{\mu}_f(S)) + \mathrm{Var}_S(\widetilde{\sigma}_f(S)).$$

Decomposition target

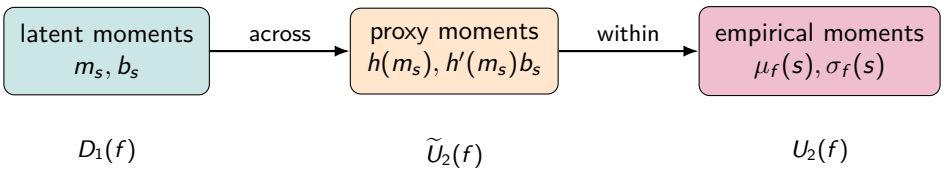
$D_1(f)$ is the leading decomposition of $\widetilde{U}_2(f)$ around the global latent mean $\bar{m}$.

Delta-method approximation: [Oehlert, 1992, van der Vaart, 1998, Serfling, 1980].

Main GLM identity

The GLM diagnostic separates the empirical criterion into a leading interpretable term and two approximation gaps:

$$U_2(f) = D_1(f) + \Delta_{\text{within}}(f) + \Delta_{\text{across}}(f)$$



- $\Delta_{\text{within}}(f) = U_2(f) - \widetilde{U}_2(f)$, measures the approximation error *within groups*, when the exact moments of $h(\eta)$ are replaced by $h(m_s)$ and $h'(m_s)b_s$.
- $\Delta_{\text{across}}(f) = \widetilde{U}_2(f) - D_1(f)$, measures the approximation error *across groups*, when $\widetilde{U}_2(f)$ is replaced by its expansion around $\bar{m}$.

Leading GLM decomposition

$$D_1(f) = \{h'(\bar{m})\}^2 \Big[\underbrace{\mathrm{Var}_S(\gamma_S)}_{\text{direct}} + \underbrace{\mathrm{Var}_S(\langle \boldsymbol{\mu}^{(S)}, \beta \rangle)}_{\text{indirect}} + \underbrace{2\,\mathrm{Cov}_S(\gamma_S, \langle \boldsymbol{\mu}^{(S)}, \beta \rangle)}_{\text{interaction}} + \underbrace{\mathrm{Var}_S(B)}_{\text{structural}} \Big]$$
$$+ \underbrace{2h'(\bar{m})h''(\bar{m})\bar{b}\ \mathrm{Cov}_S(M, B)}_{\text{curvature coupling}} + \underbrace{\{h''(\bar{m})\}^2 \bar{b}^2\, \mathrm{Var}_S(M)}_{\text{curvature amplification}}.$$

What changes from the linear case?

The four linear channels remain, rescaled by $\{h'(\bar{m})\}^2$, but the inverse link adds two curvature terms.

[Tierny et al., 2026b].

Curvature creates prediction-scale channels

Curvature coupling

$$2h'(\bar{m})h''(\bar{m})\bar{b}\ \mathrm{Cov}_S(M, B).$$

It links group differences in latent means and latent dispersions.
Its sign can be positive or negative.

Curvature amplification

$$\{h''(\bar{m})\}^2\bar{b}^2\,\mathrm{Var}_S(M).$$

It converts latent mean heterogeneity into output-scale dispersion.
It is always nonnegative.

Main message

The inverse link does not merely rescale disparities; it can create additional prediction-scale disparity channels.

Logistic regression: explicit decomposition

Here $h(\eta) = (1 + e^{-\eta})^{-1}$. Let $\bar{\mu} = h(\bar{m})$ and $a = \bar{\mu}(1 - \bar{\mu})$. Since

$$h'(t) = h(t)\{1 - h(t)\} \text{ and } h''(t) = h(t)\{1 - h(t)\}\{1 - 2h(t)\},$$

then the leading logistic decomposition is

$$D_{1,\text{logit}}(f) = a^2 \Big[\underbrace{\text{Var}_S(\gamma_S)}_{\text{direct}} + \underbrace{\text{Var}_S(c_S)}_{\text{indirect}} + \underbrace{2\,\text{Cov}_S(\gamma_S, c_S)}_{\text{interaction}} + \underbrace{\text{Var}_S(B)}_{\text{structural}} \Big]$$
$$+ \underbrace{2a^2(1 - 2\bar{\mu})\bar{b}\ \text{Cov}_S(M, B)}_{\text{curvature coupling}} + \underbrace{a^2(1 - 2\bar{\mu})^2\bar{b}^2\,\text{Var}_S(M)}_{\text{curvature amplification}},$$

where

$$c_s = \langle \boldsymbol{\mu}_s, \boldsymbol{\beta} \rangle, \qquad M = m_S, \qquad B = b_S.$$

Low-probability regime

When $\bar{\mu}$ is small, $a = \bar{\mu}(1 - \bar{\mu}) \simeq \bar{\mu}$, so the whole decomposition is compressed by roughly $\bar{\mu}^2$ on the probability scale.

Empirical reading: what to report

For each fitted model and sensitive comparison, report

$$U_2(f), \qquad \widetilde{U}_2(f), \qquad D_1(f),$$

together with

$$\Delta_{\text{within}}(f) = U_2(f) - \widetilde{U}_2(f), \qquad \Delta_{\text{across}}(f) = \widetilde{U}_2(f) - D_1(f).$$

How to read the output

- $U_2(f)$: observed output-scale disparity.
- $D_1(f)$: interpretable leading decomposition.
- gaps: reliability of the approximation.

Empirical illustration: MEPS

The paper uses MEPS-HC data and two sensitive comparisons:

Hispanic vs. non-Hispanic, uninsured vs. insured.

Model	Outcome	Scale
Linear	$\log(1 + \texttt{TOTEXP16})$	expenditure proxy
Logistic	hospitalization indicator	probability
Poisson	`OBTOTV16`	frequency
Tweedie	`OBTOTV16`	overdispersed utilization

Empirical question

Is the fitted disparity mostly due to explicit group effects, proxy covariates, dispersion structure, or nonlinear link geometry?

[Tierny et al., 2026b].

Linear regression

	Hispanic vs. non-Hispanic	**Uninsured vs. insured**
Predictive performance		
Base model: R^2	0.196 (0.004)	0.211 (0.005)
Moment-aligned model: R^2	0.178 (0.005)	0.172 (0.005)
Base model: RMSE	1.576 (0.008)	1.561 (0.008)
Moment-aligned model: RMSE	1.593 (0.008)	1.599 (0.008)
Decomposition of $U_2(f_{\text{lin}})$		
Direct mean: $\text{Var}_S(\gamma_S)$	0.035 (0.004)	0.081 (0.008)
Indirect mean: $\text{Var}_S(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)$	0.003 (0.001)	0.003 (0.001)
Interaction: $2\,\text{Cov}_S(\gamma_S, \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)$	0.019 (0.002)	0.032 (0.003)
Indirect structural: $\text{Var}_S(B)$	0.001 (0.000)	0.002 (0.000)
Total decomposed disparity	0.057 (0.006)	0.117 (0.010)

Logistic regression

	Hispanic vs. non-Hispanic	Uninsured vs. insured
Predictive performance		
Base model: deviance	11 732 (274)	11 675 (275)
Moment-aligned model: deviance	11 788 (275)	11 823 (273)
Moment-based disparity on the probability scale $(\times 10^{-4})$		
Base model: $U_2(f)$	2.34 (.520)	6.04 (.635)
Moment-aligned model: $U_2(f^{\mathrm{fair}})$	< 0.01	.034 (.022)
Components of $D_{1,\mathrm{logit}}(f)$ $(\times 10^{-4})$		
Direct mean: $a^2 \operatorname{Var}_S(\gamma_S)$	.954 (.253)	3.28 (.713)
Indirect mean: $a^2 \operatorname{Var}_S(\langle \boldsymbol{\mu}^{(S)}, \beta \rangle)$	.003 (.003)	.011 (.008)
Interaction: $2a^2 \operatorname{Cov}_S(\gamma_S, \langle \boldsymbol{\mu}^{(S)}, \beta \rangle)$	$-.078$ (.060)	.361 (.095)
Indirect structural: $a^2 \operatorname{Var}_S(B)$	.028 (.008)	.084 (.012)
Curvature coupling: $2a^2(1-2\bar{\mu})\bar{b} \operatorname{Cov}_S(M, B)$	.187 (.038)	.664 (.064)
Curvature amplification: $a^2(1-2\bar{\mu})^2\bar{b}^2 \operatorname{Var}_S(M)$	.330 (.087)	1.34 (.263)
Total leading decomposition: $D_{1,\mathrm{logit}}(f)$	1.42 (.362)	5.74 (.986)

Scope and limitations

The framework is

- a diagnostic for fitted prediction disparities;
- a decomposition of statistical channels;
- adapted to actuarial GLMs;
- useful for model audit and governance.

The framework is not

- a legal test of discrimination;
- a full causal identification strategy;
- a full characterization of distributional parity outside the linear-Gaussian benchmark;
- a normative definition of a fair tariff.

Take-away

The linear benchmark decomposes disparity into

$$\text{direct} + \text{indirect} + \text{interaction} + \text{structural}.$$

The GLM extension keeps these four channels and adds

$$+ \text{ curvature coupling} + \text{curvature amplification}.$$

Main message

In actuarial GLMs, prediction disparities are shaped not only by coefficients and covariates, but also by the inverse link transporting the latent score to the actuarial scale.

$$U_2(f) = D_1(f) + \Delta_{\text{within}}(f) + \Delta_{\text{across}}(f)$$

References

Agueh, M. and Carlier, G. (2011).
Barycenters in the wasserstein space.
SIAM Journal on Mathematical Analysis, 43(2):904–924.

Barocas, S., Hardt, M., and Narayanan, A. (2017).
Fairness in machine learning.
Nips tutorial, 1:2017.

Black, E., Yeom, S., and Fredrikson, M. (2020).
Fliptest: fairness testing via optimal transport.
In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121.

Blinder, A. S. (1973).
Wage discrimination: Reduced form and structural estimates.
The Journal of Human Resources, 8(4):436–455.

Charpentier, A. (2024).
Insurance, Biases, Discrimination and Fairness.
Springer Actuarial. Springer.

References

Charpentier, A., Flachaire, E., and Gallic, E. (2023).
Optimal transport for counterfactual estimation: A method for causal inference.
In *Optimal Transport Statistics for Economics and Related Topics*, pages 45–89. Springer.

De Jong, P. and Heller, G. Z. (2008).
Generalized linear models for insurance data.
Cambridge University Press.

De Lara, L., González-Sanz, A., Asher, N., Risser, L., and Loubes, J.-M. (2024).
Transport-based counterfactual models.
Journal of Machine Learning Research, 25(136):1–59.

Denuit, M., Hainaut, D., and Trufin, J. (2019).
Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions.
Springer.

Dowson, D. C. and Landau, B. V. (1982).
The fréchet distance between multivariate normal distributions.
Journal of Multivariate Analysis, 12(3):450–455.

References

Dunn, P. K. and Smyth, G. K. (2018).
Generalized Linear Models with Examples in R.
Springer, 2 edition.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012).
Fairness through awareness.
In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226. ACM.

Fairlie, R. W. (2005).
An extension of the blinder-oaxaca decomposition technique to logit and probit models.
Journal of economic and social measurement, 30(4):305–316.

Fernandes Machado, A., Voitsitska, I., Charpentier, A., and Gallic, E. (2026).
Sequential transport for causal mediation analysis.
arXiv, 2603.15182.

Fortin, N., Lemieux, T., and Firpo, S. (2011).
Decomposition methods in economics.
In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4A, pages 1–102. Elsevier.

References

Frees, E. W. (2009).
Regression modeling with actuarial and financial applications.
Cambridge University Press.

Fukuchi, K. and Sakuma, J. (2023).
Demographic parity constrained minimax optimal regression under linear model.
In *Advances in Neural Information Processing Systems*, volume 36, pages 8653–8689.

Gelbrich, M. (1990).
On a formula for the L^2 wasserstein metric between measures on euclidean and hilbert spaces.
Mathematische Nachrichten, 147(1):185–203.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. (2019).
Obtaining fairness using optimal transport theory.
In *International Conference on Machine Learning*, pages 2357–2365. Proceedings of Machine Learning Research.

Hardt, M., Price, E., and Srebro, N. (2016).
Equality of opportunity in supervised learning.
In *Advances in Neural Information Processing Systems*.

References

Jann, B. (2008).
The blinder–oaxaca decomposition for linear regression models.
The Stata Journal, 8(4):453–479.

Kitagawa, E. M. (1955).
Components of a difference between two rates.
Journal of the American Statistical Association, 50(272):1168–1194.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017).
Counterfactual fairness.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.

McCullagh, P. and Nelder, J. A. (1989).
Generalized Linear Models.
Chapman and Hall, 2 edition.

Oaxaca, R. (1973).
Male-female wage differentials in urban labor markets.
International Economic Review, 14(3):693–709.

References

Oehlert, G. W. (1992).
A note on the delta method.
The American Statistician, 46(1):27–29.

Ohlsson, E. and Johansson, B. (2010).
Non-life insurance pricing with generalized linear models, volume 174.
Springer.

Santambrogio, F. (2015).
Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.
Birkhäuser.

Serfling, R. J. (1980).
Approximation Theorems of Mathematical Statistics.
Wiley.

Takatsu, A. (2011).
Wasserstein geometry of Gaussian measures.
Osaka Journal of Mathematics, 48(4):1005–1026.

References

Tierny, B., Charpentier, A., and Hu, F. (2026a).
Decomposing direct and indirect biases in linear models under demographic parity constraint.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40.

Tierny, B., Charpentier, A., and Hu, F. (2026b).
Extending bias decomposition frameworks to generalized linear models.
Submitted.

van der Vaart, A. W. (1998).
Asymptotic Statistics.
Cambridge University Press.

Villani, C. (2009).
Optimal transport: old and new, volume 338.
Springer.