Does Residuals-on-Residuals Regression Produce Representative Estimates of Causal Effects?

Apoorva Lal [1] Winston Chou [2]

[1] Amazon Web Services, work done while at Netflix

[2] Netflix

August 1, 2025

Observational Causal Inference at Netflix

- We love A/B testing at Netflix.
- However, many important questions are not directly A/B testable:
  - For example, we want to know how streaming affects subscriber retention...
  - But A/B tests can only *encourage* our members to stream.
- In general, data scientists need nimble tools to explore causal questions.
- $\rightsquigarrow$ At Netflix, we maintain an internal Observational Causal Inference platform.

Residuals-on-Residuals Regression

Initially, our platform implemented residuals-on-residuals regression (RORR):

- Suppose Y_i and T_i are determined by a Partially Linear Model (PLM),

$$Y_i = \theta T_i + g(X_i) + e_i \quad \text{and} \quad T_i = h(X_i) + u_i.$$

- Estimate θ by regressing $\widetilde{Y}_i = Y_i - \widehat{g}(X_i)$ on $\widetilde{T}_i = T_i - \widehat{h}(X_i)$.

Residuals-on-Residuals Regression

Initially, our platform implemented residuals-on-residuals regression (RORR):

- Suppose Y_i and T_i are determined by a Partially Linear Model (PLM),

$$Y_i = \theta T_i + g(X_i) + e_i \quad \text{and} \quad T_i = h(X_i) + u_i.$$

- Estimate θ by regressing $\widetilde{Y}_i = Y_i - \widehat{g}(X_i)$ on $\widetilde{T}_i = T_i - \widehat{h}(X_i)$.

Pros

- Easy to explain
- Scalable to large datasets (OLS is RORR!)
- Appropriate for many questions

Cons

- Only estimates Average Treatment Effects (ATEs) if PLM is correct

Misspecification Bias of RORR for Binary Treatments

Suppose the true model is:

$$Y_i = \theta_i T_i + g(X_i) + e_i, \qquad T_i \in \{0, 1\},$$

that is, treatment effects are heterogeneous and treatment is binary.

[1] E.g., Angrist and Krueger 1999

Misspecification Bias of RORR for Binary Treatments

Suppose the true model is:

$$Y_i = \theta_i T_i + g(X_i) + e_i, \qquad T_i \in \{0, 1\},$$

that is, treatment effects are heterogeneous and treatment is binary.

The bias of $\widehat{\theta}$ relative to the ATE $E[\theta_i]$ is well understood:[1]

- Units with more variable treatment (π_i closer to $\frac{1}{2}$) receive higher weights.
- The resulting bias is proportional to the covariance between θ_i and $\pi_i(1 - \pi_i)$.
- For example, if units with $\pi \approx \frac{1}{2}$ have larger θ_i, $\widehat{\theta}$ is positively biased.

[1] E.g., Angrist and Krueger 1999

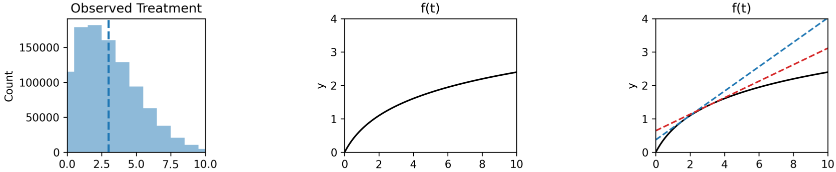
Misspecification Bias of RORR for Continuous Treatments

- What about continuous treatments?

$$Y_i = f(T_i) + g(X_i) + e_i.$$

- Two potential sources of treatment effect heterogeneity:
  1. The dose-response function $f_i(T_i)$ may be heterogeneous.
  2. Even if f_i is homogeneous, nonlinearity in f also induces heterogeneity.
- We focus on the latter.

Simple Example



In many practical applications:

- Treatments are right-skewed $\rightsquigarrow$ conditional variance of T is increasing in $E[T|X]$.
- Dose-response functions exhibit diminishing returns, so f' is decreasing in T.
- RORR is variance-weighted, skewing $\widehat{\theta}$ towards f' at larger values of T...
- ...leading to attenuation bias $E[\widehat{\theta}] < E[f'(T)]$.

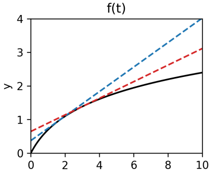
Bias Decomposition

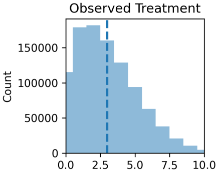
Formally, the bias of RORR can be decomposed into two parts:

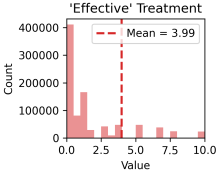
$$
\begin{aligned}
& \frac{E[(T_i - h(X_i))^2 f'(T_i^*)]}{E[(T_i - h(X_i))^2]} - E[f'(T_i)] \\
= \quad & \underbrace{\frac{E[(T_i - h(X_i))^2 f'(T_i)]}{E[(T_i - h(X_i))^2]} - E[f'(T_i)]}_{:=A} \\
+ \quad & \underbrace{\frac{E[(T_i - h(X_i))^2 f'(T_i^*)]}{E[(T_i - h(X_i))^2]} - \frac{E[(T_i - h(X_i))^2 f'(T_i)]}{E[(T_i - h(X_i))^2]}}_{:=B}
\end{aligned} \tag{1}
$$

- A is the familiar variance-weighting bias, which also appears in the binary case.
- B is unique to multi-valued treatments:
  - $\widehat{\theta}$ cannot be interpreted as a weighted average of derivatives at observed treatments.
  - Instead, it is a weighted average of derivatives at interpolated treatments.

Returning to Example



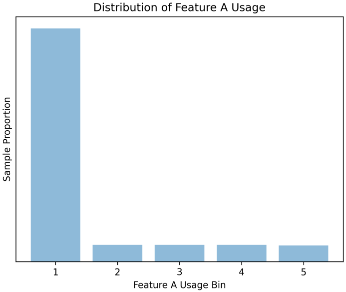




- The RORR estimand $E[\widehat{\theta}]$ is a weighted average of derivatives...
- ...evaluated on an "effective" treatment distribution that is not the observed one.

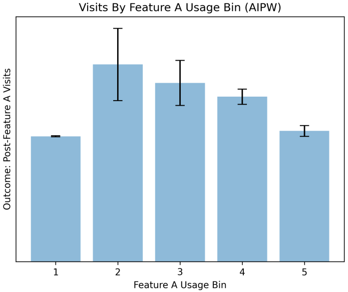
Application at Netflix

- Treatment is skewed ✓
- Dose-response function exhibits diminishing returns ✓
- $\rightsquigarrow$ RORR skews towards higher values of t, where f' is negative.



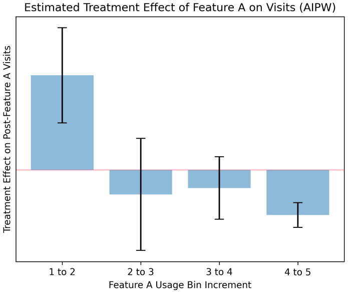
Application at Netflix

- Treatment is skewed ✓
- Dose-response function exhibits diminishing returns ✓
- $\rightsquigarrow$ RORR skews towards higher values of t, where f' is negative



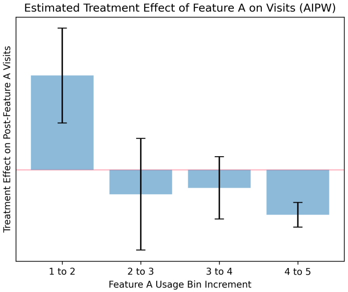
Application at Netflix

- Treatment is skewed ✓
- Dose-response function exhibits diminishing returns ✓
- $\rightsquigarrow$ RORR skews towards higher values of t, where f' is negative



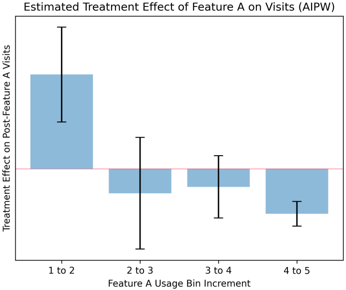
RORR Estimate is Actually Negative

RORR	Std. Err.	95% CI
-0.0038	0.001	(-0.005, -0.002)



AIPW Yields a More Representative Estimate...

RORR	Std. Err.	95% CI
-0.0038	0.001	(-0.005, -0.002)
AIPW	Std. Err.	95% CI
5.343	0.010	(5.324, 5.362)



. . . While Enabling Useful Diagnostics

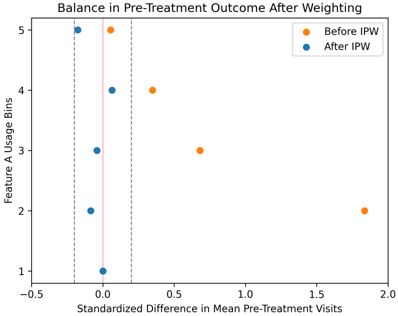


Figure: Balance in Pre-Treatment Outcomes After Inverse Propensity Score Weighting

Thanks!

Link to paper: