EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
2025, Vol. 18, Issue 4, Article Number 7150
ISSN 1307-5543 – ejpam.com
Published by New York Business Global

Sparse Cells, Inflated Odds: Bias Reduction Methods for Complex Survey Data with an Application to Hygiene Availability

Zakariya M. S. Mohammed[1,2,*], Sanaa A. Mohammed[3], Mohyaldein Salih[2], Myada A. Ibrahim[4], Omer M. A. Hamed[5], Ola A. I. Osman[5], Ali Satty[2]

[1] *Center for Scientific Research and Entrepreneurship, Northern Border University, Arar, Saudi Arabia*
[2] *Department of Mathematics, College of Science, Northern Border University, Saudi Arabia*
[3] *Department of Statistics, Faculty of Mathematical Sciences and Informatics, University of Khartoum, Sudan*
[4] *Sudan Medical Specialization Board, Khartoum, Sudan*
[5] *Department of Finance and Insurance, College of Business Administration, Northern Border University, Arar, Saudi Arabia*

Abstract. Complex survey designs, such as UNICEF's Multiple Indicator Cluster Surveys (MICS), often generate sparse cells across strata, leading to inflated odds ratios (ORs), wide confidence intervals (CIs), and convergence failures in survey-weighted logistic regression (SLR). Using Sudan MICS 2014 data on 16,679 households, this study aims to demonstrate a practical workflow to mitigate sparse-data bias: (1) SLR for population-representative estimates, (2) Firth's penalized logistic regression (FPLR) to reduce bias from sparse cells, and (3) geographic collapsing of states into broader regions to improve stability. Results highlight highly inflated ORs under SLR (e.g., OR=153.19 for Central Darfur) were substantially reduced after FPLR and regional collapsing. Substantively, strong gradients in handwashing facility access emerged by education, wealth, and geography, while sex of household head showed no effect. This study illustrates that combining design-based estimation, penalized likelihood, and collapsing strategies yields more stable inference in complex surveys with sparse data, offering practical guidance for applied researchers.

2020 Mathematics Subject Classifications: 62J12, 62D05, 62P10

Key Words and Phrases: Complex survey design, sparse data, Firth's penalized logistic regression (FPLR), survey-weighted logistic regression (SLR), multiple indicator cluster survey (MICS)

[*] Corresponding author.

DOI: https://doi.org/10.29020/nybg.ejpam.v18i4.7150

Email address: zakariya.mohammed@nbu.edu.sa (Z. M. S. Mohammed)

1. Introduction

Analysis of health and social outcomes from large-scale household surveys is central to global monitoring. Surveys such as UNICEF's Multiple Indicator Cluster Surveys (MICS) adopt stratification, clustering, and sampling weights to ensure population representativeness [1–3]. Yet these same design features often generate small cell sizes across subgroups, creating sparse data. Sparse-data bias is a well-documented challenge in logistic regression, producing inflated odds ratios (ORs), wide or infinite confidence intervals (CIs), and convergence failures [4–6]. These problems undermine valid inference, particularly when analyzing rare outcomes or highly stratified covariates.

Methodological studies show that maximum likelihood estimates under sparse data exaggerate effect sizes [7]. Penalized likelihood methods, such as Firth's logistic regression (FPLR), reduce small-sample bias and produce finite estimates under separation [8, 9]. However, integrating penalization with survey design is not straightforward because most implementations omit survey weights, clustering, and stratification [3]. Recent work advocates hybrid strategies combining survey estimation with bias-reduction methods [10]. When penalization is insufficient, collapsing sparse categories offers a practical remedy [11].

These methodological concerns are not abstract: they have critical implications for global health monitoring. Hand hygiene is among the most cost-effective interventions to reduce diarrheal disease, respiratory infections, and undernutrition, major causes of child mortality [12–14]. In Sudan, nearly one-third of households lack basic handwashing facilities, reflecting entrenched inequalities by wealth, education, and geography [15]. Sparse data in subnational survey strata complicates estimation of these inequalities, risking misleading evidence for policy.

This study aims to use Sudan MICS 2014 as a case study to examine determinants of household handwashing availability while addressing sparse-data bias. By comparing the SLR and FPLR models, and by collapsing states into broader regions, this study proposes an analytic workflow that balances representativeness, bias reduction, and stability. Recent methodological advances, such as those by [10], who conceptualized the integration of penalized estimation within survey frameworks, and [16], who extended Bayesian hierarchical approaches for sparse survey data, have provided important theoretical groundwork. Building on these developments, the present study diverges by operationalizing a practical, empirically grounded workflow that can be directly implemented with standard survey software. This workflow integrates survey-weighted logistic regression, Firth's penalized regression, and regional collapsing to stabilize estimation in real-world complex survey settings like MICS, thereby bridging the gap between emerging theory and applied implementation. The contribution is dual: providing substantive insights into hygiene inequality in Sudan, and methodological guidance for analyzing binary outcomes in complex survey contexts.

2. Method

2.1. Data source and sample

The analysis used data from the 2014 Sudan MICS, a nationally representative survey conducted by UNICEF and the Central Bureau of Statistics [17]. The survey employed a two-stage stratified sampling design covering all 18 states of Sudan, with stratification by urban and rural areas. A total of 16,801 households were interviewed, and after excluding cases with missing data on handwashing facilities or key covariates, the final analytic sample comprised 16,679 households. The outcome of interest was the availability of a designated handwashing facility with water and soap, recorded through direct observation by interviewers. Independent variables included sex and education of the household head, household wealth quintile, household size, area of residence (urban/rural), and state of residence, selected based on theoretical and empirical relevance to household sanitation and hygiene [18]. The MICS dataset was selected for its dual value: its public health importance in monitoring hand hygiene, and its methodological relevance due to complex design and sparse subnational data that challenge regression modeling.

2.2. Statistical analysis

2.2.1. Survey-weighted logistic regression (SLR).

Let $Y_i \in \{0,1\}$ indicate handwashing availability for household i, X_i be its a vector of indicator/coded covariates constructed from the study predictors (e.g. education, wealth, residence, household size, state, sex of head), and w_i is the survey weight. Define

$$p_i = P(Y_i = 1 \mid X_i), \quad \operatorname{logit}(p_i) = \eta_i = X_i^\top \beta. \tag{1}$$

The SLR estimator $\hat{\beta}$ maximizes the weighted (pseudo) log-likelihood

$$\iota_w(\beta) = \sum_i w_i \{Y_i \log p_i + (1 - Y_i) \log(1 - p_i)\}, \tag{2}$$

equivalently solving the weighted score equations

$$U_W(\beta) = \sum_i w_i X_i (Y_i - p_i) = 0, \tag{3}$$

with expected information

$$A_W(\beta) = \sum_i w_i X_i X_i^\top p_i (1 - p_i). \tag{4}$$

Under stratification and clustering, uncertainty is obtained via the design-based sandwich variance

$$\widehat{\mathrm{Var}}(\hat{\beta}) = A_W^{-1}\, \beta_W\, A_W^{-1}, \tag{5}$$

where β_W aggregates cluster-level score residuals within strata. ORs for any contrast c are $\exp\{c^\top(\hat{\beta})\}$ with Wald CIs: $\exp\{c^\top(\hat{\beta}) \mp z_{1-\alpha/2}\sqrt{c^\top\widehat{\mathrm{Var}}(\hat{\beta})c}\}$. When some design cells are small, $p_i(1-p_i)$ can be near zero and A_W becomes ill-conditioned, yielding inflated ORs, very wide (even infinite) CIs, and occasional non-convergence (quasi/complete separation).

2.2.2. Firth's penalized logistic regression (FPLR).

Let $\iota(\beta)$ be the standard (unweighted) log-likelihood. Firth's estimator $\hat{\beta}_F$ maximizes the penalized log-likelihood

$$\iota^*(\beta) = \iota(\beta) + 0.5\log|\mathcal{I}(\beta)|, \tag{6}$$

where the usual (unweighted) log-likelihood

$$\iota(\beta) = \sum_i \{Y_i \log p_i + (1-Y_i)\log(1-p_i)\}, \tag{7}$$

and the expected Fisher information is

$$\mathcal{I}(\beta) = \sum_i X_i X_i^\top p_i(1-p_i). \tag{8}$$

Equivalently, $\hat{\beta}_F$ solves the penalized score equations

$$U^*(\beta) = \frac{d\iota^*(\beta)}{d\beta} = \sum_i X_i(Y_i - p_i) + 0.5\frac{d}{d\beta}\log|\mathcal{I}(\beta)| = 0, \tag{9}$$

which are implemented via modified Iteratively Reweighted Least Squares (IRLS) and yield finite, bias-reduced estimates under separation and sparse cells. The penalization is based on the Jeffreys prior. Inference is obtained from penalized profile likelihood intervals for linear contrasts $c^\top(\hat{\beta})$; ORs are: $\exp\{c^\top(\hat{\beta}_F)\}$ (Wald intervals may also be formed using the observed penalized information at $\hat{\beta}_F$). Note that standard Firth's implementations are unweighted; thus, while FPLR stabilizes estimates under sparsity, it does not fully integrate survey weights, clustering, or stratification.

2.2.3. Collapsing and regrouping (post-SLR and Firth).

Now, to reduce sparsity, let each household i belong to one of the 18 states encoded by a one-vector $d_i \in \{0,1\}^{18}$. Define a many-to-one mapping matrix $M \in \{0,1\}^{18}$ (with $K < 18$) that groups states into broader regions (e.g., Northern, Eastern, Western, Central, Khartoum). The regional indicator for household i is

$$z_i = Md_i \in \{0,1\}^K. \tag{10}$$

Constructing the covariate vector X_i by replacing state indicators with z_i, the logistic model retains the same linear predictor but with regional rather than state effects:

$$P(Y_i = 1 \mid X_i) = p_i, \quad \text{logit}(p_i) = X_i^\top \beta. \tag{11}$$

Collapsing increases the effective weighted cell counts in each geographic category,

$$N_k^{\text{eff}} = \sum_{i:z_{ik}=1} w_i, \tag{12}$$

which improves conditioning of the (expected) information matrix, design-based for SLR,

$$A_W(\beta) = \sum_i w_i X_i X_i^\top p_i(1 - p_i), \tag{13}$$

or unweighted for Firth,

$$\mathcal{I}(\beta) = \sum_i X_i X_i^\top p_i(1 - p_i). \tag{14}$$

Intuitively, by increasing the least-populated regional cell, one can mitigate sparsity; this lowers $\text{Var}(c^\top \hat{\beta})$ yields a narrower CIs for the ORs, $\exp\{c^\top(\hat{\beta})\}$.

2.2.4. Software and implementation.

Analyses were performed in Stata (version 17). The survey design was declared with svyset using the household pweight, strata, and PSU variables. Design-adjusted descriptive statistics and Rao–Scott Chi-square tests were obtained via svy: prefix commands [19]. The primary model was SLR using svy: logit, reporting design-based standard errors (linearized) and 95% CIs. Bias reduction under sparsity was examined with Firth's logistic regression (firthlogit) [20]. In keeping with current software constraints for complex designs, FPLR was estimated without survey weighting and treated as robustness checks to assess inflation and instability observed in SLR. To mitigate sparsity, states were collapsed into broader regions; models were re-fit with regional indicators under both SLR and FPLR. Cases missing the outcome or key predictors were excluded, yielding a final analytic sample of 16,679 households. Statistical significance was set at $\alpha = 0.05$.

3. Results

3.1. Household and regional correlates.

Table 1 summarizes household characteristics, handwashing facility availability across categories, and Chi-square test results for their associations. The results indicated that handwashing facility availability did not differ by sex of the household head (41% in male-headed households vs. 42% in female-headed, $p = 0.452$). Strong socioeconomic and geographic gradients were observed. Only 34% of households headed by individuals with no education had a handwashing facility, compared with 66% among those with higher education ($p < 0.005$). Availability increased across wealth quintiles, from 23% in the poorest households to 71% in the richest ($p < 0.005$). Urban households reported greater access than rural ones (47% vs. 38%, $p < 0.005$). Marked spatial disparities were evident, ranging from 2% in Gadarif and 8% in East Darfur to 93% in Central Darfur and over 60% in North Darfur and South Kordofan ($p < 0.005$). Household size showed smaller differences, with 37% availability among households of 1–3 members versus 42% in larger households ($p < 0.005$). This descriptive evidence provides a foundation for subsequent regression analyses by pinpointing where inequalities in hygiene access are most pronounced.

Table 1: Household characteristics and rational correlates.

Variables	Category	Household		Observed place for hand washing		Chi-square	p-value
		Count	%	Count	%		
Sex of household head	Male	14414	86%	5882	41%	0.566	0.452
	Female	2387	14%	994	42%		
Education of household head	None	7799	46%	2657	34%	592.397	< 0.001
	Primary	4730	28%	1854	39%		
	Secondary	3137	19%	1650	53%		
	Higher	1013	6%	666	66%		
	Missing	122	1%	NA	NA		
Area	Urban	5000	30%	2365	47%	119.756	< 0.001
	Rural	11801	70%	4510	38%		
State	Northern	423	3%	203	48%	3490.814	< 0.001
	River Nile	666	4%	430	65%		
	Red Sea	519	3%	82	16%		
	Kassala	722	4%	143	20%		
	Gadarif	858	5%	17	2%		
	Khartoum	2317	14%	1228	53%		
	Gezira	2629	16%	1508	57%		
	White Nile	874	5%	435	50%		
	Sinnar	661	4%	356	54%		
	Blue Nile	656	4%	343	52%		
	North Kordofan	1125	7%	40	4%		
	South Kordofan	462	3%	285	62%		
	West Kordofan	1003	6%	228	23%		
	North Darfor	1243	7%	776	62%		
	West Darfor	553	3%	261	47%		
	South Darfor	1282	8%	220	17%		
	Central Darfor	299	2%	279	93%		
	East Darfor	508	3%	42	8%		
Number of household members	1–3	3343	20%	1242	37%	25.201	< 0.001
	4–6	7037	42%	2924	42%		
	7 & more	6420	38%	2710	42%		
Wealth index quintile	Poorest	3368	20%	784	23%	1861.534	< 0.001
	Second	3592	21%	1099	31%		
	Middle	3339	20%	1258	38%		
	Fourth	3209	19%	1390	43%		
	Richest	3293	20%	2344	71%		

3.2. SLR versus FPR: Resolving sparse-data instability.

Table 2 compares results from SLR and FPLR, focusing on the problem of inflated ORs caused by sparse data. Compared with SLR, FPLR consistently reduced extreme

ORs in sparse states, often by 20–80%. For example, Central Darfur's estimate declined from 153.2 (SLR) to 143.0 (FPLR), while River Nile fell from 20.0 to 7.5. Similarly, Khartoum dropped from 12.4 to 2.8. These reductions confirm that penalization shrinks inflated estimates toward more plausible values while narrowing CIs. Education and wealth gradients showed a similar pattern: for higher education, the OR fell from 3.71 (SLR) to 1.37 (FPLR), illustrating that much of the inflation observed in the survey model was corrected. Similarly, River Nile decreases from 20.00 to 7.54, and Khartoum from 12.41 to 2.80. These results illustrate how penalization narrows CIs and produces more stable estimates, particularly in categories with limited observations. Importantly, while FPLR reduces extreme values, some odds ratios remain high (e.g., North Darfur OR=22.91, South Kordofan OR=10.03), reflecting persistent sparse-data challenges. Education and wealth variables also demonstrate this pattern. For instance, the "Higher education" category shows a sharp reduction: OR=3.71 in the unadjusted survey model, but only 1.37 (95% CI: 1.15–1.64) with Firth's penalization. Similarly, the richest wealth quintile decreases from OR=8.14 (survey) to OR=11.74 (95% CI: 9.64–14.32) in the FPLR model, still indicating strong inequality but with more controlled inflation. Overall, these findings confirm that FPLR reduces inflation and improves precision, but because residual instability persists, the next step involved collapsing and regrouping states into broader regions to further stabilize the estimates.

Table 2: SLR with FPLR to address inflated ORs.

Variables	Category	SLR						FPLR		
		Unadjusted Prevalence OR (UPOR)			Adjusted Prevalence OR (APOR)			Adjusted Prevalence OR (APOR)		
		OR	95% CI		OR	95% CI		OR	95% CI	
			Lower	Upper		Lower	Upper		Lower	Upper
Education of household head										
	Primary	1.25***	1.09	1.43	0.91***	0.80	1.05	0.96	0.87	1.05
	Secondary	2.15***	1.78	2.59	1.10***	0.92	1.30	1.14***	1.02	1.27
	Higher	3.71***	2.82	4.87	1.48*	1.14	1.93	1.37*	1.15	1.64
	None	1			1			1		
Area										
	Urban	1.45*	1.08	1.95	0.88	0.62	1.23	0.909	0.82	1.008
	Rural	1			1			1		
State										
	Northern	10.14***	5.40	19.05	2.33*	1.17	4.65	2.69***	2.02	3.60
	River Nile	20.00***	10.40	38.46	6.19***	3.20	11.98	7.54***	5.69	10.07
	Red Sea	2.07*	1.02	4.23	0.91	0.51	1.63	0.81	0.60	1.10
	Kassala	2.72***	1.39	5.33	1.59	0.88	2.86	1.57***	1.18	2.11
	Gadarif	0.22***	0.10	0.50	0.14***	0.06	0.29	0.16***	0.09	0.25
	Khartoum	12.41***	6.67	23.12	2.76***	1.50	5.08	2.80***	2.12	3.72
	Gezira	14.80***	8.00	27.39	5.20***	2.71	9.99	3.99***	3.03	5.31
	White Nile	10.90***	5.71	20.78	5.56***	2.90	10.66	5.27***	4.02	6.97
	Sinnar	12.88***	7.47	22.23	6.34***	3.56	11.30	5.91***	4.51	7.81
	Blue Nile	12.08***	6.76	21.59	7.18***	3.83	13.46	5.89***	4.51	7.77
	North Kordofan	0.40*	0.18	0.88	0.31***	0.16	0.60	0.32***	0.21	0.48
	South Kordofan	17.66***	9.86	31.64	14.28***	7.84	25.99	10.02***	7.71	13.15
	West Kordofan	3.24***	1.91	5.50	3.42***	2.05	5.71	3.69***	2.81	4.88
	North Darfor	18.25***	10.25	32.49	23.81***	13.76	41.21	22.90***	17.57	30.16
	West Darfor	9.79***	5.28	18.13	10.26***	5.30	19.88	8.56***	6.58	11.22
	South Darfor	2.28**	1.26	4.14	2.23***	1.31	3.83	2.20***	1.66	2.93
	Central Darfor	153.19***	84.22	278.65	172.76***	92.87	321.37	143.03***	102.37	202.98
	East Darfor	1			1			1		
Number of household members										
	1–3	0.81***	0.70	0.93	0.91	0.78	1.07	0.95	0.86	1.05
	4–6	0.97	0.89	1.06	1.00	0.90	1.12	0.99	0.91	1.08
	7 & more	1			1			1		
Wealth index quintile										
	Second	1.45***	1.16	1.83	1.87***	1.45	2.40	1.61***	1.41	1.84
	Middle	1.99***	1.47	2.70	3.22***	2.28	4.56	2.71***	2.33	3.15
	Fourth	2.52***	1.76	3.60	3.86***	2.55	5.84	3.05**	2.57	3.63
	Richest	8.14***	5.90	11.22	14.98***	9.60	23.37	11.73***	9.63	14.31
	Poorest	1			1			1		

** Significant at 0.05, ** significant at 0.01, *** significant at 0.001*

3.3. SLR versus FPLR: Collapsing states into regions.

Table 3 reports the results of SLR and FPLR after applying a collapsing and regrouping strategy to combine states into broader regional categories. This re-categorization was designed to mitigate sparse-data problems identified earlier and to produce more stable and conservative estimates. The results confirm that collapsing states into regions reduces instability and yields narrower CIs, while the main associations remain consistent in direction and significance. For household head's education, the FPLR estimates are smaller but remain significant for higher education (Firth's OR=1.23; 95% CI: 1.05–1.46), confirming that education is positively associated with availability of handwashing facilities. For primary and secondary education, the ORs are closer to 1 in FPLR, suggesting that much of the inflation seen in the survey model has been addressed. Urban–rural differences follow the same pattern: while the survey model shows an inflated OR (OR=1.45), FPLR produces a conservative estimate with an OR below 1 (Firth's OR=0.80; 95% CI: 0.73–0.88), indicating lower adjusted odds for handwashing facilities in urban households after controlling for covariates. The regrouping strategy for states is particularly effective. The inflated ORs observed in the earlier model (Table 2) are now reduced to more stable regional estimates. For example, Khartoum falls from OR=2.30 (survey) to Firth's OR=0.57 (95% CI: 0.48–0.68), and Northern Sudan from OR=2.83 to Firth's OR=0.83 (95% CI: 0.72–0.96). Eastern Sudan shows a consistently protective effect, with the OR further stabilized under penalization (Firth's OR=0.14; 95% CI: 0.12–0.15). These findings demonstrate how collapsing states into broader regions reduces sparse-data bias and produces more interpretable results. Wealth-related inequalities remain strong and robust across both models. For instance, households in the richest quintile show more than nine times higher odds of having a handwashing place compared to the poorest (Firth's OR=9.88; 95% CI: 8.30–11.77). The gradient across wealth quintiles persists, with consistent increases from the second through richest groups, highlighting persistent socioeconomic disparities. In summary, the collapsing and regrouping of states into regions improved model stability and yielded more conservative estimates, yet the associations between handwashing availability and key predictors, education, region, area of residence, and wealth, remain significant and directionally consistent.

Table 3: SLR versus FPLR after collapsing states into regions.

Variables	Category	Unadjusted Prevalence OR (UPOR)			Adjusted Prevalence OR (APOR) – SLR			Adjusted Prevalence OR (APOR) – FPLR		
		OR	Lower	Upper	OR	Lower	Upper	OR	Lower	Upper
Education of household head										
	Primary	1.25***	1.09	1.43	0.87*	0.77	0.99	0.80***	0.74	0.87
	Secondary	2.15***	1.78	2.59	1.06	0.90	1.26	0.97	0.87	1.07
	Higher	3.71***	2.82	4.87	1.45**	1.13	1.87	1.23*	1.04	1.45
	None	1			1			1		
Area										
	Urban	1.45*	1.08	1.95	0.81	0.58	1.13	0.79***	0.72	0.87
	Rural	1			1			1		
Region										
	Northern Sudan	2.82***	1.95	4.08	0.98	0.63	1.51	0.83***	0.72	0.95
	Eastern Sudan	0.26***	0.18	0.38	0.17***	0.12	0.23	0.13*	0.11	0.15
	Khartoum	2.30***	1.43	3.68	0.67	0.42	1.08	0.57***	0.48	0.67
	Central Sudan	2.47***	1.83	3.34	1.25	0.92	1.69	0.88***	0.79	0.97
	Western Sudan	1			1			1		
Number of household members										
	1–3	0.81***	0.70	0.93	0.86	0.74	1.00	0.94	0.85	1.03
	4–6	0.97	0.89	1.06	0.98	0.88	1.08	0.99	0.92	1.07
	7 & more	1			1			1		
Wealth index quintile										
	Second	1.45***	1.16	1.83	1.64***	1.27	2.09	1.86***	1.67	2.06
	Middle	1.99***	1.47	2.70	2.19***	1.56	3.08	2.46***	2.17	2.78
	Fourth	2.52***	1.76	3.60	2.59***	1.70	3.92	2.62***	2.26	3.04
	Richest	8.14***	5.90	11.22	9.78***	6.29	15.22	9.87***	8.30	11.77
	Poorest	1			1			1		

** Significant at 0.05, ** significant at 0.01, *** significant at 0.001*

4. Discussion

This study addressed the persistent challenge of sparse-data bias in complex surveys, using Sudan MICS 2014 as a case study on household handwashing availability. Conventional survey-weighted logistic regression produced inflated ORs and wide CIs, particularly for small state-level strata, confirming earlier warnings about instability under quasi- or complete separation [4, 21]. Applying FPLR reduced these artifacts by shrinking extreme estimates toward plausible values, consistent with its established role in bias reduction for rare events [5, 8, 22]. Yet residual instability persisted in very sparse strata, demonstrating that penalization alone cannot fully resolve separation when coupled with complex design features. Collapsing states into broader regions provided an additional remedy, narrowing confidence intervals while preserving effect directions. Taken together, the results illustrate a feasible workflow, design-based regression for representativeness, penalized regression for bias reduction, and collapsing for stability, that can guide analysts working with MICS and other large household surveys [10].

Substantively, the findings highlight entrenched inequalities in access to hygiene facilities across Sudan. Education and wealth gradients remained strong even after methodological corrections, echoing global evidence that socioeconomic status is a robust determinant of sanitation access [14, 18]. Geographic disparities were also pronounced, with extremely low coverage in states such as Gadarif compared with high access in Central Darfur, consistent with spatial inequities documented in water, sanitation, and hygiene (WASH) research in sub-Saharan Africa [23]. These results suggest that expanding hygiene access requires both pro-poor investments and regionally targeted interventions, as recommended by [15]. Importantly, the methodological refinements applied here did not erase substantive patterns of inequality but instead produced more credible estimates, ensuring that policy responses are guided by stable and interpretable evidence.

This study has several strengths, including the use of a nationally representative dataset, the systematic comparison of analytic strategies, and the transparent demonstration of sparse-data bias and remedies. Nonetheless, limitations must be noted. Current implementations of Firth's regression do not fully incorporate survey weights or clustering [3], which may leave residual bias. This limitation has motivated emerging methodological efforts toward weighted penalized likelihood and pseudo-likelihood formulations that combine survey weights with penalization to achieve design-consistent inference [24]. Collapsing the 18 states into five broader regions improved model stability and effective sample size but may have obscured within-region heterogeneity, a trade-off commonly observed in sparse-data analyses [11]. However, this regional grouping was guided by both statistical and substantive considerations, as the merged states share similar socioeconomic and health profiles, allowing stability without sacrificing interpretability within Sudan's administrative structure. Finally, as a cross-sectional survey, MICS does not allow causal inference, and unmeasured confounders, such as water reliability or cultural practices, may partly explain observed associations. Future research should explore Bayesian hierarchical and penalized multilevel models that can simultaneously accommodate survey design and small-area estimation [16]. By combining current methodological discussion with substantive interpretation, this study contributes both to improving analytic practice in complex survey research and to strengthening the evidence base for achieving Sustainable Development Goal 6 on universal hygiene access.

5. Conclusion

This study shows that sparse cells in complex surveys like MICS can distort design-based logistic regression, yielding inflated ORs, wide CIs, and occasional non-convergence. Using a hygiene-access case study, the study implemented a practical analytic strategy: estimate SLR for population-representative effects; apply FPLR to reduce separation-driven bias and ensure finite, more stable coefficients; and, when sparsity persists, collapse small geographic units into broader regions to further stabilize estimates. Substantively, access to handwashing facilities varied little by sex of household head but showed strong gradients by education, wealth, residence, and geography. Because sparse-data bias is structural in complex survey designs, it must be addressed to produce credible, policy-relevant inference. In our MICS case study, SLR alone produced unstable, inflated estimates in sparse strata; FPLR reduced this inflation, and regrouping states into broader regions improved precision with modest loss of granularity. Together, these steps yield stable, interpretable estimates that strengthen the evidence base for targeting hygiene investments toward poorer households, less-educated heads, rural areas, and lagging regions. Practically, this approach offers a robust template for analyzing binary outcomes in complex surveys while retaining policy-relevant signals of inequality in hygiene access. The adoption of bias-adjusted workflows, such as the one demonstrated here, can strengthen evidence-based policy formulation and better guide investments toward communities most at risk of being left behind in achieving Sustainable Development Goal 6.

Ethical considerations

This study used secondary data from the publicly available 2014 Sudan MICS, accessible via the UNICEF MICS database (https://mics.unicef.org/surveys). All data are fully anonymized, and no identifiable information was used. As the analysis involved publicly available secondary data, no additional ethical approval was required.

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number “NBU-FFR-2025-1635-08”.

References

[1] S G Heeringa, B T West, and P A Berglund. *Applied Survey Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2017.

[2] E L Korn and B I Graubard. *Analysis of Health Surveys*. Wiley, 1999.

[3] T Lumley. *Complex surveys: A guide to analysis using R*. John Wiley & Sons, Hoboken, NJ, 2010.

[4] G King and L Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001.

[5] G Heinze and M Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 2002.

[6] S Nemes, J M Jonasson, A Genell, and G Steineck. Bias in odds ratios by logistic regression modeling and sample size. *BMC Medical Research Methodology*, 9:56, 2009.

[7] S Greenland, M A Mansournia, and D G Altman. Sparse data bias: A problem hiding in plain sight. *BMJ*, 352:i1981, 2016.

[8] D Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[9] R Puhr, G Heinze, M Nold, L Lusa, and A Geroldinger. Firth’s logistic regression

with rare events: Accurate effect estimates and predictions? *Statistics in Medicine*, 36(14):2302–2317, 2017.

[10] I Burstyn and D Miller. Adjustment for sparse data bias in odds ratios: Significance to appraisal of risk of diabetes due to occupational trichlorfon insecticide exposure. *Global Epidemiology*, 8:100154, 2024.

[11] S Dörken, M Avalos, E Lagarde, and M Schumacher. Penalized logistic regression with low prevalence exposures beyond high-dimensional settings. *PLoS ONE*, 14(5):e0217057, 2019.

[12] V Curtis and S Cairncross. Effect of washing hands with soap on diarrhoea risk in the community: A systematic review. *The Lancet Infectious Diseases*, 3(5):275–281, 2003.

[13] M C Freeman, M E Stocks, O Cumming, A Jeandron, J P T Higgins, J Wolf, and V Curtis. Hygiene and health: Systematic review of handwashing practices worldwide and update of health effects. *Tropical Medicine & International Health*, 19(8):906–916, 2014.

[14] J Wolf, R Johnston, M C Freeman, P K Ram, T Slaymaker, E Laurenz, and A Prüss-Ustün. Handwashing with soap after potential faecal contact: Global, regional and country estimates. *International Journal of Epidemiology*, 47(4):1204–1218, 2018.

[15] World Health Organization (WHO) and UNICEF. Progress on household drinking water, sanitation and hygiene 2000–2020: Five years into the sdgs. Technical report, World Health Organization and UNICEF, Geneva, 2021.

[16] M Stolte. A comprehensive review of bias reduction methods for logistic regression. *Statistical Surveys*, 2024.

[17] UNICEF Sudan and Central Bureau of Statistics. Sudan multiple indicator cluster survey 2014, final report. Technical report, UNICEF, Khartoum, Sudan, 2016.

[18] A Prüss-Ustün, J Wolf, J Bartram, T Clasen, O Cumming, M C Freeman, and R Johnston. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on low- and middle-income countries. *International Journal of Hygiene and Environmental Health*, 222(5):765–777, 2019.

[19] StataCorp. *Stata 19 Survey Data Reference Manual.* Stata Press, College Station, TX, 2025.

[20] J Coveney. firthlogit: Stata module to calculate bias reduction in logistic regression, 2008. Statistical Software Components S456948, Boston College Department of Economics.

[21] E Vittinghoff and C E McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, 165(6):710–718, 2007.

[22] S Greenland and M A Mansournia. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, 34(23):3133–3143, 2015.

[23] G D Demissie, Y Yeshaw, W Aleminew, and K Shitu. Diarrhea and associated factors among under five children in sub-saharan africa: Evidence from demographic and health surveys of 34 sub-saharan countries. *PLOS ONE*, 16(9):e0257522, 2021.

[24] A Iparragirre, D Lee, B Sarr, and P Congdon. Variable selection with lasso regression for complex survey data. *Stat*, 12(1):e578, 2023.