

Asian Journal of Probability and Statistics

Volume 28, Issue 5, Page 46-60, 2026; Article no.AJPAS.156473
ISSN: 2582-0230

Multilevel Beta Regression Model for Potato Post-Harvest Losses along the Farm-Market Value Chain in Kenya

Erick Kirui [a*], Ayubu Anapapa [a] and John Mutuguta [a]

[a] Murang'a University of Technology, Nairobi, Kenya.

Authors' contributions

This work was carried out in collaboration among all authors. Author EK was the main author and contributed to the conception, design and implementation of the study. Authors AA and JM provided guidance and supervision on research writing, methodology and analysis. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/ajpas/2026/v28i5894

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://pr.sdiarticle5.com/review-history/156473

Original Research Article

Received: 03/02/2026
Accepted: 20/04/2026
Published: 25/04/2026

Abstract

Post-harvest losses in potato value chain is one of the major threats to food security and farmers livelihoods in Kenya. This paper uses multilevel beta regression modeling approach to study the determinants and spatial heterogeneity of post-harvest potato losses in Kenya farm-market value chain using data from stratified cross-sectional survey of 628 value chain actors within Nakuru and Nyandarua counties, the major potato producing highlands in the country. Five models were estimated in succession; Poisson baseline, single-level Negative Binomial (NB), multilevel NB, single-level Beta regression, and the final multilevel Beta regression model. We found that the multilevel beta regression is the best model (AIC=2711.30, log-likelihood=1375.65, LRT: $\chi^2 = 27.07, \mathrm{df} = 2, \mathrm{p} < 0.001$) confirming appropriateness of Beta distribution for the bounded proportional loss and need for multilevel model. The storage methods were the most powerful explanatory variable ($Wald\ \chi^2 = 1182.71, p < 0.001$); the cold storage reduce proportional loss by 13.07 percentage points loss while dark room storage by 7.44 percent relative to the traditional storage methods. Storage duration, market

**Corresponding author: Email: echeruiyotkirui@gmail.com;*

Cite as: *Kirui, E., Anapapa, A., & Mutuguta, J. (2026). Multilevel Beta Regression Model for Potato Post-Harvest Losses along the Farm-Market Value Chain in Kenya. Asian Journal of Probability and Statistics, 28(5), 46–60. https://doi.org/10.9734/ajpas/2026/v28i5894*

distance, the usage of chemicals, extension and training and transportation method also proved to be highly significant explanatory variables while the education of actor had insignificant impact on potato post-harvest losses. Random intercepts at county level are close to zero and likelihood ratio test indicated a significant model fit improvement with multilevel model. Residual diagnostics by DHARMa showed validity of Beta distribution assumption. The findings are policy-relevant in investment decisions aimed at reduction of post-harvest losses along the Kenyan potato value chain through provision of access to cheaper storage infrastructure, integration of extension-input system and transport infrastructure improvement.

Keywords: Multilevel beta regression; post-harvest losses; negative binomial.

1 Background of the Study

Agriculture is the staple of economic growth in most developing nations especially in sub-Saharan Africa. The industry is a significant contributor to gross domestic product (GDP) and has high percentage workforce in rural regions and supports household food security and income source. The Food and Agriculture Organization of the United Nation's reports that agriculture contributes around 23% of GDP in sub-Saharan Africa and provides more than 60% of labor force (Olakiumide, 2021). Although there have been tremendous improvements in agricultural production and policy support, there is still a high percentage of food produce that goes to waste before it reaches final consumers. These losses, which are also referred to as post-harvest-losses (PHL), are also one of the most urgent problems that food systems in developing economies face (Gill & Sharma, 2021).

Post-harvest losses refer to quantifiable quantitative and qualitative losses in food quantity, quality, or nutritional value that takes place between point of harvest and the point of consumption. This loss is experienced at different points of the agricultural value chain such as during harvesting, handling, storage, processing, transportation and marketing. It has been estimated that developing countries suffer post-harvest losses of between 20 and 40 percent of all agricultural production each year (Olakiumide, 2021), (Zsögön et al., 2022). In the case of perishable commodities like fruits, vegetables and root crops, these numbers may be even greater. The impacts are extensive, including loss of income in smallholder farmers, high consumer prices, food insecurity and unwarranted environmental impacts due to the wastages of land, water and energy (Mburu et al., 2023; Mugo et al., 2021).

The potato crop in Kenya forms a very essential part of the national food system as both staple food and a very important source of income to millions of smallholder farming households. The crop is mainly grown in highlands of the country such as the Rift Valley, Central and East, that offer the cool weather and fertile soils that support growth of tubers (Mwakidoshi, 2021). Current statistics show that Kenya yields about 2.5 million metric tons of potato every year, which makes it one of the top potato producing nations in the sub-Saharan region (Mburu et al., 2023).

The post-harvest loss data are difficult to statistically analyze in several aspects. Post-harvest losses are often given as proportions or rates i.e. the percentage of total produce that is lost at a particular stage and these measures lie between zero and one. These proportional results are not in line with the assumptions of the traditional linear regression models, which assume the presence of a continuous unbounded response variable with constant error variance (Chen & Yang, 2024). Beta regression was specifically written to regress continuous proportional observations on the open interval (0, 1) using a logit link function, and is capable of dealing with the skewness and heteroscedasticity inherent in proportional data (Geissinger et al., 2022).

This research paper addresses these methodological challenges by developing and applying multilevel beta regression model to analyze determinants and spatial heterogeneity of post-harvest potato losses across the potato farm-market value chain in Kenya. It is a very powerful methodology that has an excellent theoretical basis applied to approximate the determinants of proportional losses, quantify the between-county, and between-stage variation, and generate policy-relevant information regarding how to minimize post-harvest losses in the Kenyan potato industry by reconciling the flexibility of the beta distribution plus the hierarchical nature of multilevel modeling. The objective of the study was to compare the performance of multilevel Poisson, Negative Binomial and Beta regression models in the analysis of determinants of post-harvest losses along the potato farm-market value chain in Kenya.

2 Literature Review

Post-harvest losses (PHL) are the quantifiable both in quantity and quality losses in agricultural produce between the time of harvest and the time of final consumption. The concept includes not only physical losses the weight or volume of produce decreases but also the qualitative losses that decrease the nutritional value, marketability, or consumer safety (Levaj et al., 2023). According to the estimates made by Food Agricultural Organization (2019), about a third of all food that is produced worldwide to be consumed by humans is lost or wasted, with post-harvest losses within this category making a significant part of this data (Mugo et al., 2021; Olakiumide, 2021). In sub-Saharan Africa, the total losses in form of cereals at the farm and storage levels have been estimated to cost the smallholder farmers a total of US 4 billion each year (Ariong et al., 2023; Meheretu et al., 2022).

2.1 Poisson Regression in Agricultural and Post-Harvest Loss Modeling

The Poisson regression is the model of the analysis of count data and has been used extensively in the agricultural sciences to model discrete count non-negativity outcomes, including the number of pests, the occurrence of a disease and the number of units of produce lost during storage or handling (Ma et al., 2025). The model is built on the assumption that the response variable is a Poisson variable with mean equal to the variance and a log linking function was used so that it is the expected count that is related to a linear predictor of covariates (Saha et al., 2026). Poisson regression has been applied to food loss studies to estimate the quantity of damaged units at each stage of the supply chain, which serves as a statistically based baseline model.

2.2 Negative Binomial Regression

The most common extension of the Poisson regression to the over dispersed count data is negative binomial (NB) regression model. It is defined as a Poisson-Gamma mixture, but the Poisson rate parameter itself is a Gamma-distributed random variable, with an extra dispersion parameter, which loosens the equidispersion condition (Morris & Sellers, 2022). The NB model has found broad application in health sciences, ecology, and agricultural economics to model counts with more variability than is possible with the Poisson model. Negative binomial 2 parameterization also abbreviated $NB(\mu, \theta)$ when $var(Y) = \mu + \frac{\mu^2}{\theta}$ is by far the most used in practice since it can have dispersion that increases with the mean, which is empirically true of most agricultural and ecological count data (Almasi et al., 2016; Khan & Hasan, 2022).

2.3 Beta Regression for Proportional Outcomes

Ferrari and Cribari-Neto (2004) introduced beta regression to the statistical literature as a regression model specifically designed to estimate continuous response variables that are in the open unit interval $(0, 1)$ (Chen & Yang, 2024). The beta distribution reparametrized using a mean parameter μ and a precision parameter ϕ and related the mean to a linear predictor using the logit function. Smithson and Verkuilen (2006) demonstrated that the ordinary least squares (OLS) with proportion data gives biased coefficient estimates, inappropriate standard errors and predicted values that do not lie under the $(0, 1)$ interval (Abonazel et al., 2022). Beta regression has been applied to model various proportional results in the agriculture, economics and health sciences (Marina Pare, 2025). They demonstrated that the ordinary least squares (OLS) with proportion data gives biased coefficient estimates, inappropriate standard errors and predicted values that do not lie under the $(0, 1)$ interval that are removed when using beta regression.

3 Materials and Methods

3.1 Sampling Design and Data Collection

Data was collected in Nakuru and Nyandarua counties, which are among the most significant potato growing areas in Kenya. Both of the counties are located in the central highlands with cool weather, fertile volcanic soil, and consistent rainfalls that are very suitable in growth of potatoes. Nakuru County is a place that is highly productive in agriculture and is also inhabited by high density of wholesale and retail markets along the Nakuru Nairobi path. Nyandarua County is a county in the Aberdare Ranges and is among the highest potato-producing

counties in the country and contributes a significant percentage of potatoes being consumed in Nairobi and the other urban areas (Olakiumide, 2021). A stratified sampling technique was used since it divides a population into small groups called strata based on geographical regions where potatoes are produced. A sample from each stratum was selected, and the sample size was determined using Cochran's formula for population proportions: $n = \frac{z^2 p(1-p)}{e^2}$, Where n is the sample size z is the Z-score which is based on the confidence level p is the estimated population and e the error margin. A questionnaire survey was carried out to identify distinct factors associated with post-harvest losses, planting time, quality of seeds, harvesting time, insects' infestation, and climatic conditions. The questionnaire reliability was tested using Cronbach's Alpha ($\alpha \geq 0.7$) and validity through the content validity index (CVI). R statistical software was used to analyze the data.

3.2 Study Design

This study adopted a cross-sectional analytical design involving farmers and retailers. The research was carried out in two counties in Kenya that were purposely selected, which were considered important in the production and marketing of potatoes in the country. Let Y_{ijk} represent the number of lost units of potatoes by individual i $(i = 1,2,\dots,n)$ in county $j, (j = 1,2)$ at the stage of the value chain number $(k = 1,2,\dots,k)$. The overall number of respondents were 628 distributed between Nakuru ($j = 1;\ n = 388$) and Nyandarua ($j = 2;\ n = 240$). The observations within each county are further subdivided into each value chain stage which is where a respondent is majorly operating.

3.3 Multilevel Negative Binomial Model

Post harvest count data were modeled using Negative binomial model relaxes the equi-dispersion assumption by introducing a dispersion parameter θ. $Y_{ijk} \sim NB(\mu_{ijk}, \theta)$

$$log(\mu_{ijk}) = \beta_0 + \sum_{r=1}^{p} \beta_r X_{ijk} + \mu_j + v_k \quad (1)$$

Where $\mu_j \sim N(0, \sigma_u)$ is the random effects for county $j = 1,2$ and $v_k \sim N(0, \sigma_v)$ is the random effects for stage k.

3.3.1 Likelihood Function

Let $f(Y_{ijk} \mathrm{I} \mu_{ijk}, \theta)$ be the negative binomial probability function. The conditional likelihood is $L_c = \prod_{j=1}^{2} \prod_{k=1}^{k} \prod_{i=1}^{n} f(Y_{ijk} \mathrm{I} \mu_{ijk}, \theta)$. The marginal likelihood integrates over random effects so that:

$$L = \prod_{k=1}^{K} \int \prod_{j=1}^{2} \prod_{i=1}^{n} f(Y_{ijk} \mathrm{I} \mu_k, v_k) g(\mu_k) g(v_{jk}) dv_{jk} d\mu_k \quad (2)$$

where $f(Y_{ijk} \mathrm{I} \mu_k, v_k)$ is conditional negative binomial density and $g(\mu_k), g(v_{jk})$ represent normal densities of a random effects where $\mu_k \sim N(0, {\sigma_\mu}^2)$ and $v_{jk} \sim N\left(0, {\sigma_{v_{jk}}}^2\right)$. The marginal likelihood integrates over random effects and is approximated using Laplace transformation which expands the log-likelihood around the maximum using a second-order Taylor-expansion. The significance of fixed effects is assessed using the Wald test statistic $Z = \frac{\hat{\beta}}{SE(\hat{\beta})}$, $SE(\hat{\beta})$ and random effects using the likelihood ratio test.

3.3.2 Interclass Correlation Coefficients (ICC)

This measures the proportion of the total variation in the response that is attributable to clustering effects. The county $ICC_{County} = \frac{{\sigma_\mu}^2}{{\sigma_\mu}^2 + {\sigma_v}^2 + {\sigma_e}^2}$ where ${\sigma_\mu}^2$is the county-level variance, ${\sigma_v}^2$ is stage-level variance and ${\sigma_e}^2$ is the residual variance. The stage-level ICC: $ICC_{stage} = \frac{{\sigma_v}^2}{{\sigma_\mu}^2 + {\sigma_v}^2 + {\sigma_e}^2}$. an ICC greater than 0.05 is considered sufficient to warrant a multilevel approach.

3.4 Multilevel Beta Regression Model

Let P_{ijk} denote the proportional loss incurred by individual i in county j at the value chain stage k where $P_{ijk} \in (0,1)$. Under the beta regression parameterization suggested by (Ferrari & Cribari-Neto, 2004) such that $P_{ijk} \sim Beta(u_{ijk}, \emptyset)$ where u_{ijk} is the conditional mean also called the proportional loss and $\emptyset > 0$ is the precision parameter. The conditional mean is linked to predictors via logit function given by: $logit(\mu_{ijk}) = log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \eta_{ijk}$ and the multilevel beta regression model is given by:

$$logit(\mu_{ijk}) = \beta_0 + \sum_{i=1}^{p} \beta_r X_{ijk} + \mu_j + w_k \quad (3)$$

where $\mu_j \sim N(0, {\sigma_u}^2)$, $w_k \sim N(0, {\sigma_w}^2)$ are county and stage level random effects.

4 Results and Discussions

4.1 Model Selection and Fit Statistics

Five models were estimated sequentially. Poisson baseline model generated a deviance dispersion ratio of 0.244 and a $Pearson's\ \chi^2\ /\ degree\ of\ freedom\ ratio\ \left(\frac{\chi^2}{df}\right)$ of 0.243 ($p = 1.00$) indicating that no overdispersion was found. The NB dispersion parameter deviated to infinity ($\theta = 2{,}393{,}074$) showing the NB family is not better at the count outcome than Poisson, and the likelihood ratio test was non-significant ($\chi^2 = -0.006, p = 1.000$). In the case where the loss is proportional, the single-level Beta regression was at pseudo- Pearson $R^2 = 0.7612$. The multilevel Beta model was even a better fit ($log - likelihood$ 1375.65 vs 1362.12), and the LRT was very significant ($\chi^2 = 27.07, df = 2, p < 0.001$) as shown in Table 1 below.

Table 1. Model comparison statistics

Model	Log-Likelihood	AIC	BIC	Parameters
Poisson (SL)	-1718.41	3470.82	3546.18	17
NB (SL)	-1718.41	3472.83	3552.62	17
NB (ML) *	NA	NA	NA	20
Beta Regression (SL)	1362.12	-2688.23	-2608.44	21
Multilevel Beta Regression ✓	**1375.65**	**-2711.30**	**-2622.64**	**20**

Note: Negative binomial Hessian not positive definite; SL and ML denote single level and multilevel; highlighted row indicates selected model

4.2 Intraclass Correlation Coefficients (ICC)

The null multilevel Beta model on the logit scale was computed to generate intraclass correlation coefficients (ICC) using the recommended latent Level-1 residual variance of ${\sigma_\varepsilon}^2 = \frac{\pi^2}{3} = 3.290$ as recommended by (Devine et al., 2024). The variance at the county level was ${\sigma_u}^2 = 0.000817$ ($ICC = 0.02\%$) and the variance at the stage level was ${\sigma_v}^2 = 0.001342$ ($ICC = 0.04\%$). Though the estimated values are lower than conventional 0.05, this represent a negligible proportion of the total variation of high-level grouping units. The likelihood ratio test (LRT), which was very significant ($\chi^2 = 27.07, df = 2, p = 0.001$), and proves the multilevel structure to be a significant factor that enhances the model fit. ICC estimates are in Table 2 below.

Table 2. Intraclass correlation coefficient

Grouping Level	Variance (σ^2)	ICC	ICC (%)	Justified?
County (j = 1, 2)	0.000817	0.0002	0.02%	NO
Value Chain Stage (k = 1, 2)	0.001342	0.0004	0.04%	NO
Level-1 residual variance ($\sigma^2\varepsilon = \pi^2/3$)	3.2899	—	—	—
LRT: Multilevel vs Single-level Beta	$\chi^2 = 27.07$	$df = 2$	$p < 0.001$	**YES ✓**

4.3 Poisson Regression Results

The count-based loss outcome was estimated using the Poisson regression as a baseline. The model explained significant variation ($null\ deviance = 591.91, residual\ deviance = 147.80, df = 605$). All predictor effects showed consistent direction and significance values with those of the multilevel Beta regression, which cross-model validated the fixed effect structure. Fig. 1 shows the IRR forest plot of the multilevel NB model, which graphically summarizes the predictors that significantly increase or decrease the incidence rate of count-based losses. The most significant protective and risk-increasing effects are in cold storage ($IRR = 0.649$) and in dark room storage ($IRR = 0.798$), and in storage over one month ($IRR = 1.116$) and market distances greater than 10 km ($IRR = 1.111$) respectively. Full estimates are in Table 3.

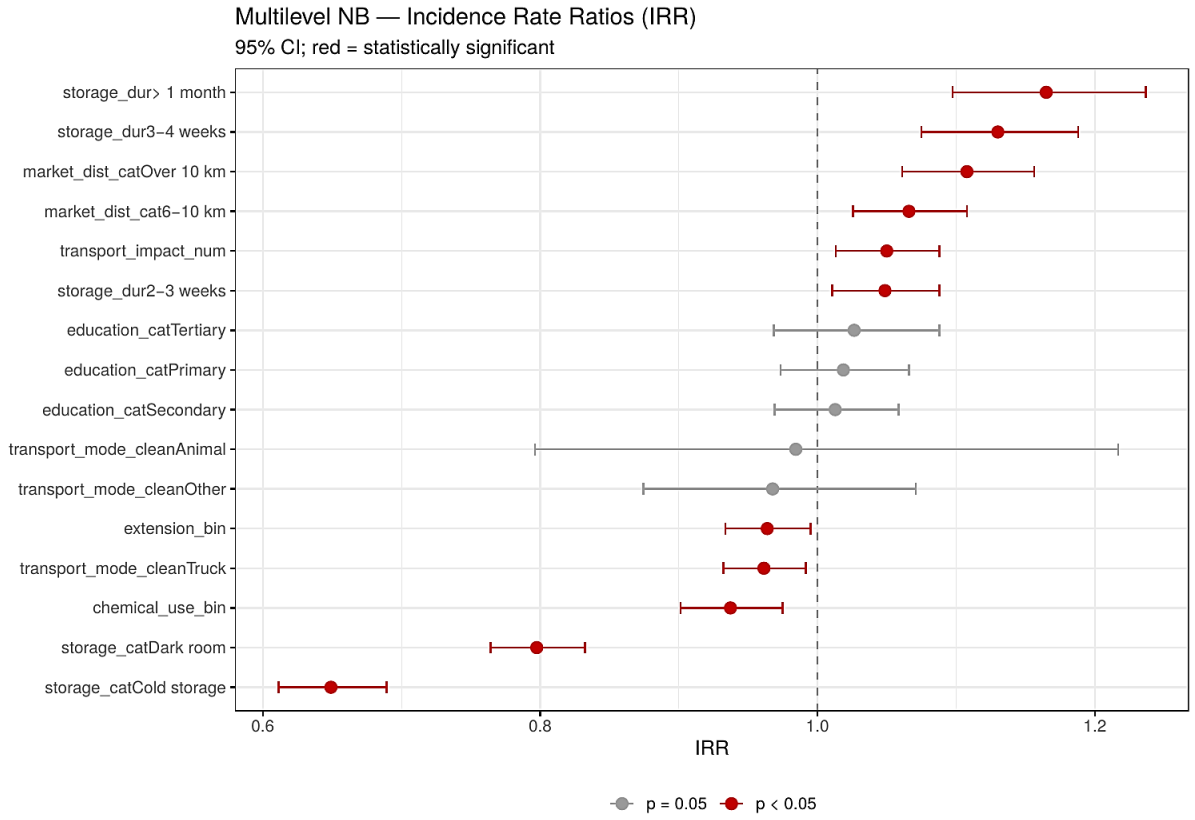


Fig. 1. Multilevel negative binomial model

Fig. 1 gives Multilevel Negative Binomial model that is; Incidence Rate Ratios (IRR) with 95 percent confidence interval. Markers with red color denote significant predictors (p value less than 0.05). Reference categories: Traditional storage, less than 2 weeks storage, less than 5 km market distance, no education, Motorcycle transport.

Table 3. Poisson fixed random effects

Variable	β	SE	z	p-value	Sig.	IRR
(Intercept)	3.344	0.058	57.38	< 0.001	***	28.31
Storage: Dark room	-0.218	0.020	-10.72	< 0.001	***	0.804
Storage: Cold storage	-0.429	0.030	-14.12	< 0.001	***	0.651
Duration: 2–3 weeks	0.052	0.018	2.90	0.004	**	1.053
Duration: 3–4 weeks	0.130	0.024	5.36	< 0.001	***	1.139
Duration: > 1 month	0.161	0.029	5.51	< 0.001	***	1.174
Chemical use (Yes)	-0.060	0.019	-3.11	0.002	**	0.942

Variable	β	SE	z	p-value	Sig.	IRR
Extension training (Yes)	-0.033	0.015	-2.16	0.031	*	0.967
Market distance: 6–10 km	0.065	0.020	3.30	0.001	***	1.067
Market distance: > 10 km	0.106	0.022	4.86	< 0.001	***	1.112
Transport impact (severity)	0.048	0.018	2.69	0.007	**	1.049
Education: Primary	0.015	0.023	0.67	0.506	—	1.016
Education: Secondary	0.012	0.022	0.55	0.586	—	1.012
Education: Tertiary	0.026	0.030	0.89	0.376	—	1.027
Transport mode: Truck	-0.041	0.015	-2.64	0.008	**	0.960
Transport mode: Animal	-0.030	0.108	-0.28	0.781	—	0.971
Transport mode: Other	-0.037	0.052	-0.71	0.476	—	0.964

The Incidence ratio rate $IRR = exp(\beta)$, The reference categories are; the traditional storage method, storage duration less than 2 weeks, market distance less than 5km, no education, mode of transport is motorcycle.

4.4 Multilevel Beta Regression Results

Multilevel Beta regression model was estimated using the value chain stage and county as the random intercept grouping factors. The model converged ($AIC = -2711.30, BIC = -2622.64, log-likelihood = 1375.65, 2$, precision parameter $\phi = 308$). The high precision parameter implies that the proportional loss distribution is narrowly clustered around the mean, which is consistent with the narrow range of losses observed (15% to 45 percent). The Smithson and Verkuilen compression $P^* = \frac{P(n-1)+0.5)}{n}$ was used to transform proportional losses to ensure values strictly lie in $(0, 1)$. Fixed effects were all analyzed based on odds ratios ($OR = exp(B)$) and marginal effects in percentage points when the sample mean was 0.319 (31.9%), which was the mean of the sample. Fig. 2 shows the odds ratio forest plot of the multilevel Beta model, which gives a visual representation of the direction and significance of all the predictors on the proportional loss outcome. Full estimates are in Table 4.

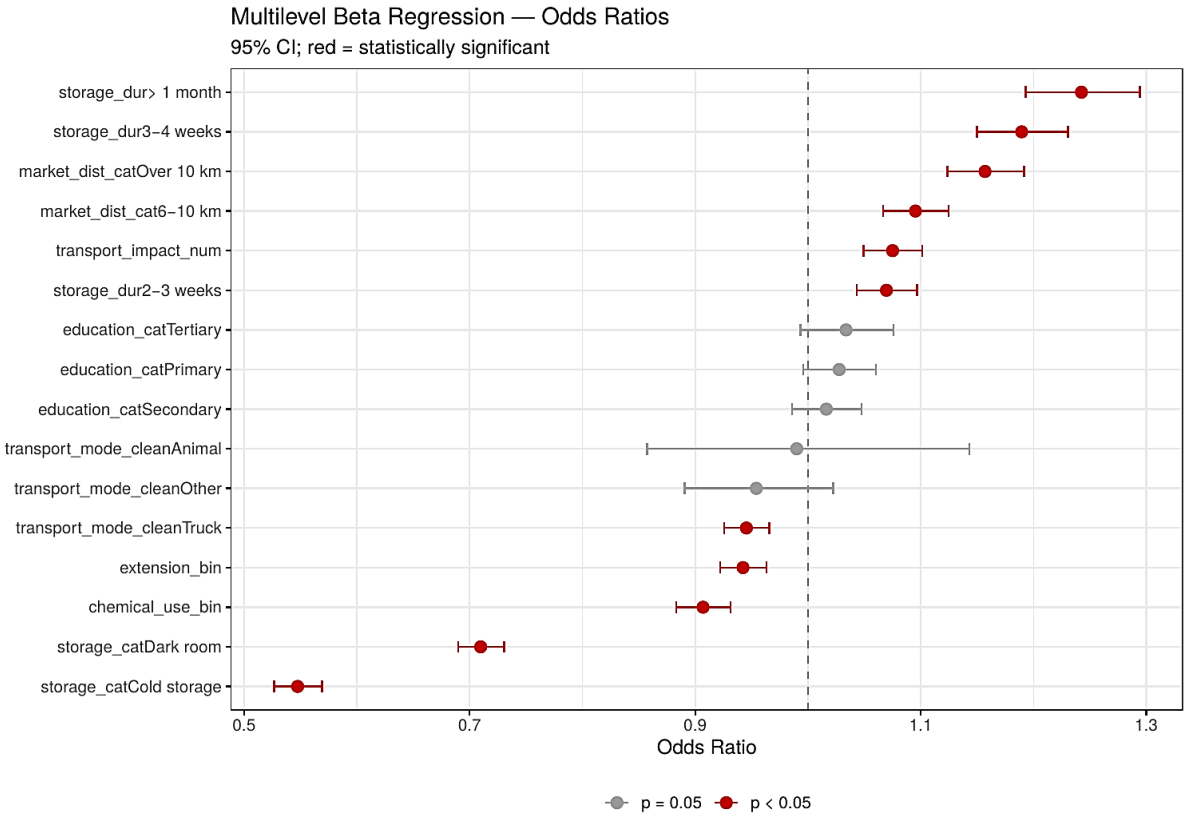


Fig. 2. Multilevel beta regression

Fig. 2 Gives the Multilevel Beta Regression summaries: Odds ratios with 95 percent confidence intervals. Red markers represent statistically significant predictors ($p < 0.05$). $OR < 1$ means a corresponding proportional decrease in the loss; $OR > 1$ means corresponding proportional increase in the loss.

4.5 Independent Variables

4.5.1 Storage Method and Storage Duration

The single most significant predictor was storage method ($Wald\ \chi^2 = 1182.71, df = 2, p = 0.001$) . Compared to traditional storage, dark room storage decreased the log-odds of proportional loss by $\beta = -0.343$ ($OR = 0.710, ME = -7.44pp, p = 0.001$) and cold storage by $\beta = -0.602$ ($OR = 0.548, ME = -13.07pp$, $p = 0.001$). Fig. 3 indicates the proportional loss by method of storage likely to be the model and clearly demonstrates the gradual assumption of 30% under the traditional storage method to 25% under the dark room and 20% under the cold storage method control of all other predictors. Storage time also had a monotonically increasing effect on losses ($Wald\ \chi^2 = 137.39, df = 3, p < 0.001$): $+1.46pp$ at 23 weeks, and $+4.71pp$ after more than one month as shown in Fig. 4. The use of chemicals was found to reduce the number of losses ($\beta = -0.098, OR = 0.907, ME = -2.12\ pp, Wald\ \chi^2 = 51.93, p = 0.001$), and users were predicted to lose about 29% as compared to 33% of non-users who did not use the chemicals.

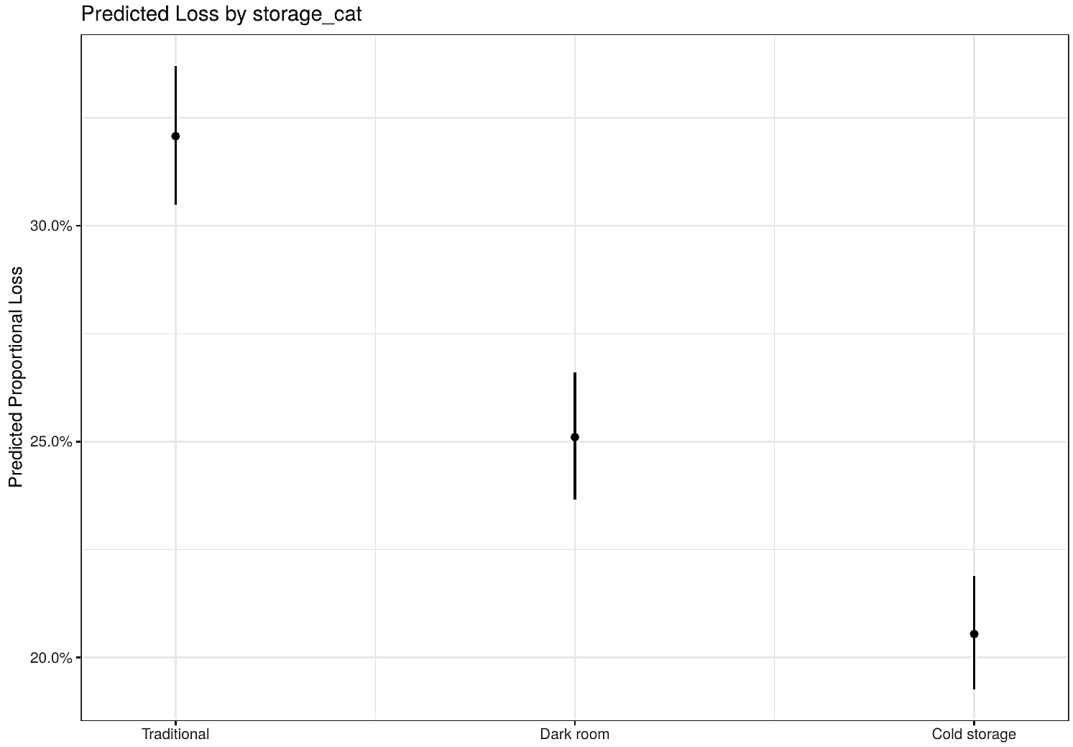


Fig. 3. Proportional loss by method of storage

Fig. 3 predicted proportional loss to storage method adjusted on all other covariates. Error bars denote confidence intervals of 95% intervals

Table 4. Multilevel beta regression coefficient random effects

Variable	β	SE	z	p-value	Sig.	OR	ME (pp)
(Intercept)	-0.908	0.052	-17.53	< 0.001	***	0.403	—
Storage: Dark room	-0.343	0.015	-23.41	< 0.001	***	0.710	-7.44
Storage: Cold storage	-0.602	0.020	-30.31	< 0.001	***	0.548	-13.07
Duration: 2–3 weeks	0.067	0.013	5.26	< 0.001	***	1.070	+1.46
Duration: 3–4 weeks	0.174	0.017	10.03	< 0.001	***	1.190	+3.77

Variable	β	SE	z	p-value	Sig.	OR	ME (pp)
Duration: > 1 month	0.217	0.021	10.45	< 0.001	***	1.240	+4.71
Chemical use (Yes)	-0.098	0.014	-7.21	< 0.001	***	0.907	-2.12
Extension training (Yes)	-0.059	0.011	-5.36	< 0.001	***	0.942	-1.29
Market distance: 6–10 km	0.091	0.013	6.78	< 0.001	***	1.100	+1.98
Market distance: > 10 km	0.146	0.015	9.73	< 0.001	***	1.160	+3.17
Transport impact (severity)	0.072	0.012	5.84	< 0.001	***	1.080	+1.57
Education: Primary	0.027	0.016	1.71	0.087	.	1.030	+0.59
Education: Secondary	0.016	0.016	1.04	0.298	—	1.020	+0.35
Education: Tertiary	0.033	0.020	1.63	0.102	—	1.030	+0.72
Transport mode: Truck	-0.056	0.011	-5.20	< 0.001	***	0.945	-1.22
Transport mode: Animal	-0.010	0.073	-0.14	0.891	—	0.990	-0.22
Transport mode: Other	-0.047	0.035	-1.33	0.183	—	0.954	-1.02

The Odds ratio $OR = exp(\beta)$, $ME\ (pp)$ are the marginal effects in percentage points at $\mu = 0.319$ (31.9%). *The reference categories: Storage duration less than 2 weeks, distance less than 5km, mode of transport is motorcycle.*

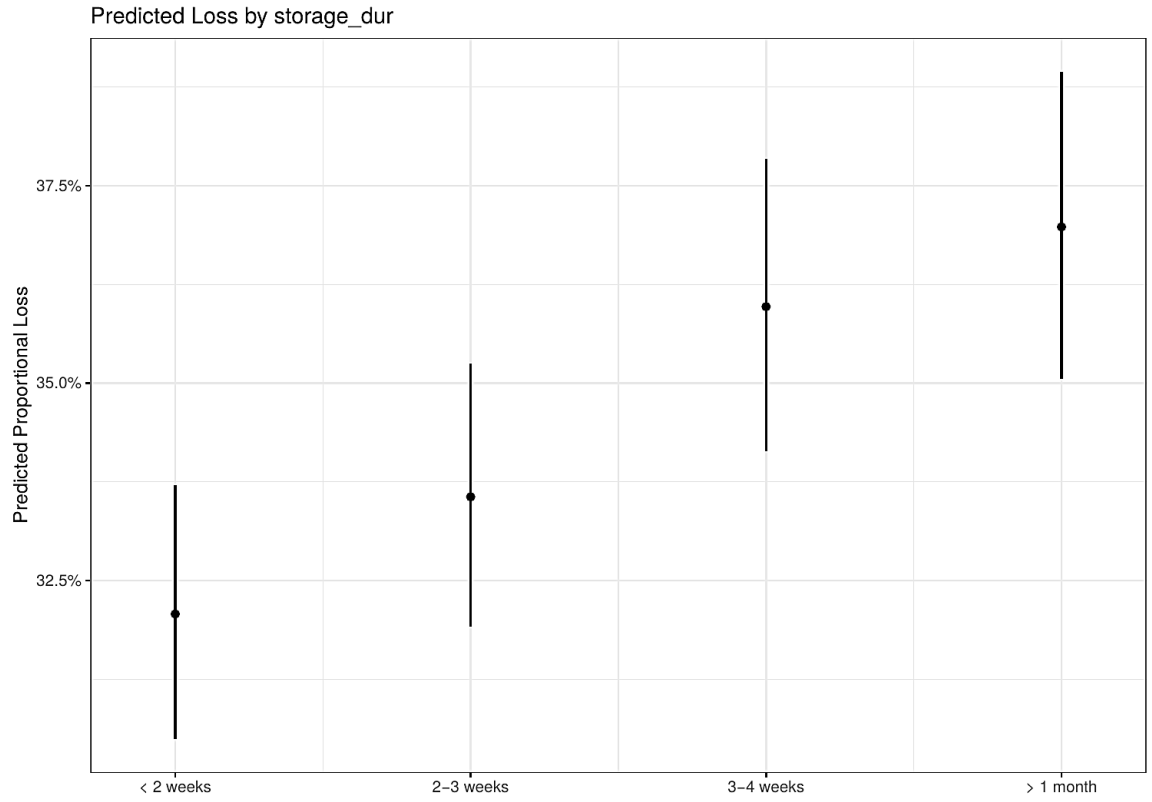


Fig. 4. Proportional loss by Duration of storage

Predicted proportional loss by length of storage predicted by other covariates, other covariates held constant. The gradual, unvarying growth is an indication of deterioration of the body.

4.5.2 Extension Services and Education Level

Proportional losses were also reduced significantly by extension training ($\beta = -0.059, OR = 0.942, ME = -1.29pp, Wald\ \chi^2 = 28.68, p = 0.001$), and trained respondents were estimated to lose about 31% versus 35% in the case of an untrained respondent. In contrast, the level of formal education was not a statistically significant predictor ($Wald\ \chi^2 = 4.06, 3, p = 0.255$) . The predicted education-based losses indicate a relatively linear trend in education variables where the predicted proportional losses vary within a tight range of 31 to 34 % once the variables of operation are held constant across the education groups. This is in contrast to the unadjusted means where the relationship was non-linear and it is therefore confirmed that the apparent

education effect in the raw data is explained by the confounding effects of operational variables like storage method and chemical use rather than by a direct effect of education on loss behavior.

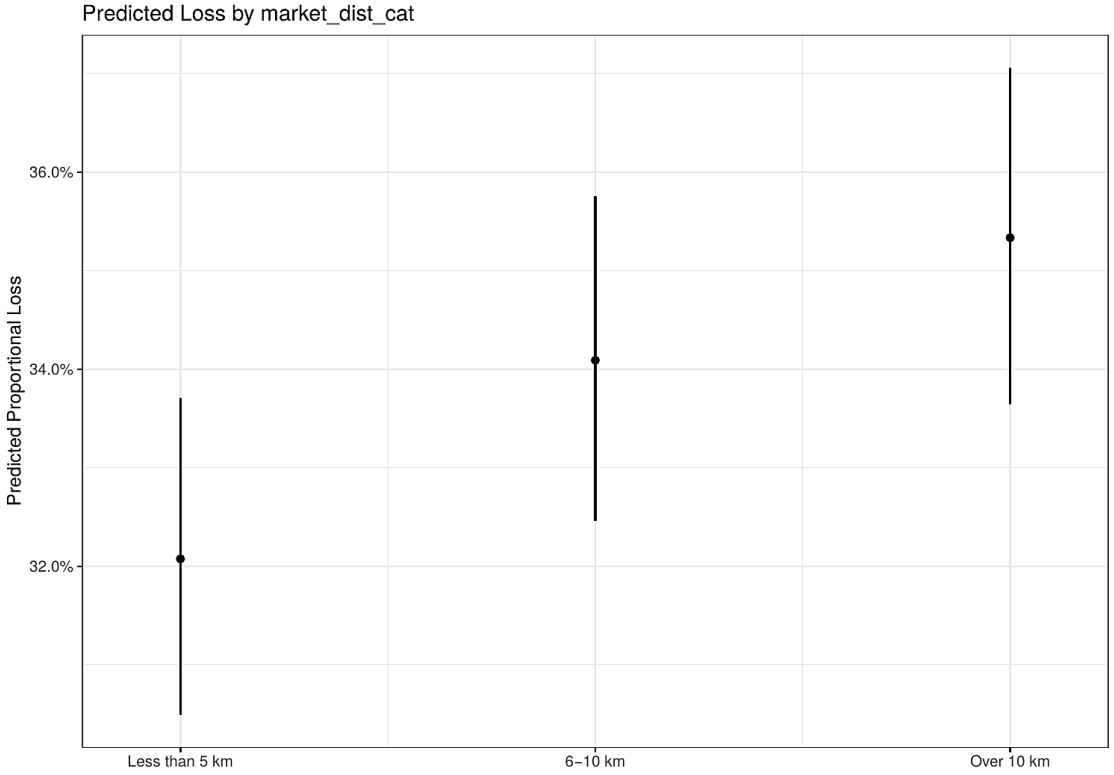


Fig. 5. Proportional loss modeled by market distance

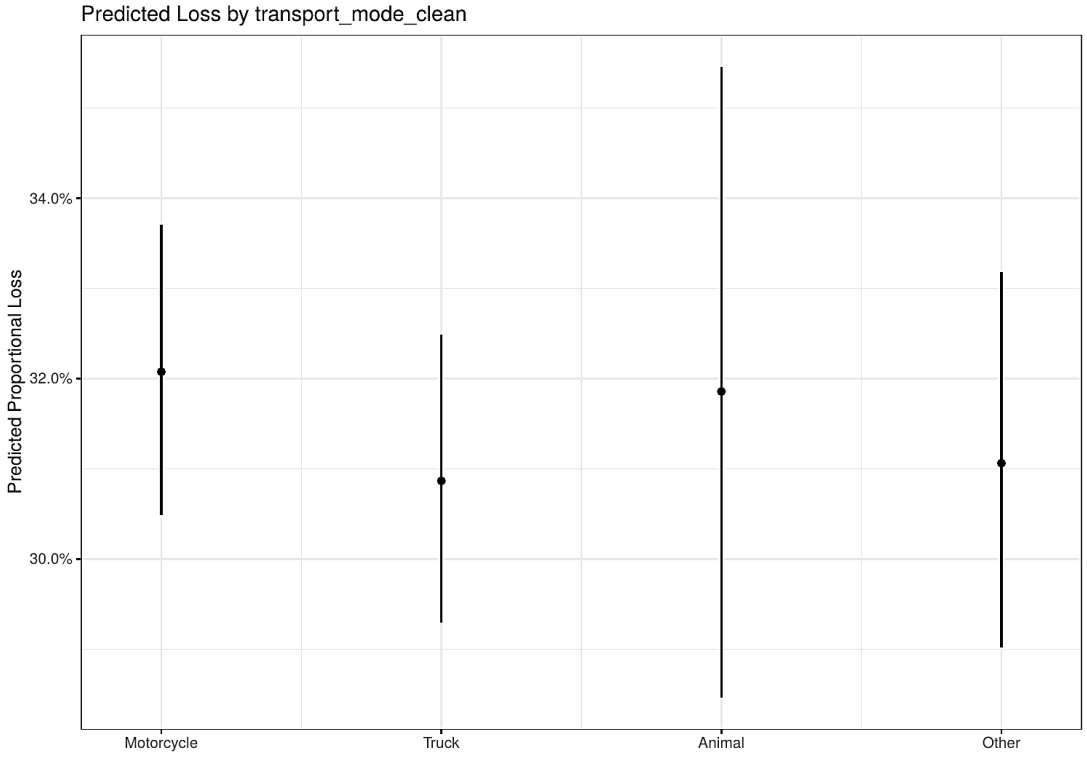


Fig. 6. Model-predicted proportional loss by mode of transportation

4.5.3 Market Access and Mode of Transportation

Market distance was an extremely important positive predictor ($Wald\ \chi^2 = 95.97, df = 2, p = 0.001$). Compared to those within 5 km of market, those within 6-10 km experienced $+1.98pp$ ($\beta = 0.091, OR = 1.10, p < 0.001$), and those outside the market beyond 10 km experienced $+3.17pp$ ($\beta = 0.146, OR = 1.160, p < 0.001$). Fig. 5 shows model-predicted losses by market distance, and it is clear that there is a step increase in the percentage of approximately 32% to 36% across the three distance categories. Perceived severity of transport also significantly increased losses by $1.57pp\ per\ unit$ ($Wald\ \chi^2 = 34.10, p < 0.001$). Overall transport mode was also important ($Wald\ \chi^2 = 27.46, df = 3, p = 0.001$) as the mode of transport that generated a $1.22pp$ reduction in comparison to motorcycle (Fig. 6). There was no significance in animal and other modes with their very small samples ($n = 3\ and\ n = 14$ respectively).

4.6 Wald Tests

Type III Wald Chi-square tests compared the overall significance of each predictor at all levels of predictor described by one or more dummies. Operational and market access variables were very significant. Storage method had the largest statistic ($\chi^2 = 1182.71, df = 2, p = 0.001$), then storage duration ($\chi^2 = 137.39$), market distance ($\chi^2 = 95.97$), chemical use ($\chi^2 = 51.93$), transport impact ($\chi^2 = 34.10$), extension training ($\chi^2 = 28.68$), and transport mode ($\chi^2 = 27.46$). Education level was not a significant predictor ($\chi^2 = 4.06, 3\ df\ =\ 0.255$). The results are shown in Table 5.

Table 5. Type III chi-square tests. multilevel beta regression

Predictor	Wald χ^2	df	p-value	Significance
Storage method	1182.71	2	< 0.001	*** Highly significant
Storage duration	137.39	3	< 0.001	*** Highly significant
Chemical use	51.93	1	< 0.001	*** Highly significant
Extension training	28.68	1	< 0.001	*** Highly significant
Market distance	95.97	2	< 0.001	*** Highly significant
Transport impact	34.10	1	< 0.001	*** Highly significant
Transport mode	27.46	3	< 0.001	*** Highly significant
Education level	4.06	3	0.255	Not significant

4.6.1 Model Diagnostics

The multilevel Negative Binomial and multilevel Beta regression models were subjected to $DHARMa$ simulation-based residual diagnostics to determine the appropriateness of their distributional assumptions. Fig. 7 and Fig. 8 shows the $DHARMa$ diagnostic plots of each model respectively. In the multilevel NB model (Fig. 7), the QQ plot of residuals shows a clear S-shaped deviation to the diagonal reference line; the KS-test ($p = 0$) and the dispersion-test ($p = 0$) are both statistically significant. The plot of residual versus fitted indicates systematic curvature, and the overall adjusted quantile test is also significant. These findings indicate that the multilevel NB model cannot be used on this dataset because overdispersion is not present in the count-based outcome ($Pearson's\ \chi^2/df = 0.243$) and the NB dispersion parameter theta approaches infinity. This gives a solid empirical validation that the NB family does not fit the count-based loss outcome in this research and the Poisson and multilevel Beta regression specifications are the right modeling options.

Fig. 7 shows $DHARMa$ residual diagnostics Multilevel Negative Binomial model. Left panel: QQ plot of simulated residuals with a large deviation in the expected uniform distribution (KS test $p = 0.006$, dispersion test $p = 0.612$ not significant, outlier test= 0.189 not significant). Right panel shows the results of the residual versus model prediction (rank transformed) with systematic curvature and significant sum test of quantile. These two panels confirm model misspecification, which is also in line with the lack of overdispersion in the count outcome and the departure of the NB dispersion parameter toward infinity ($\theta \rightarrow 2{,}393{,}074$).

The high deviations in the multilevel NB diagnostics plots are not surprising considering the nature of the data, the count-based loss outcome is not over dispersed ($Pearson's\ \chi^2/df = 0.243$) and the NB dispersion parameter was brought to non-zero in the estimation process, effectively reducing the NB to a Poisson. The

DHARMa residual diagnostics of the multilevel Negative Binomial model shows obvious model misspecification. The QQ plot indicates a systematic departure of the distribution and the Kolmogorov-Smirnov (KS) test and the dispersion test are both significant. Further, the residual-versus-fitted plot indicates that there are organized trends, implying that the model assumptions are violated. These findings verify that the Negative Binomial model cannot fit the data as it does not have overdispersion and the dispersion parameter tends to infinity. These findings also justify the need to use proportional beta regression framework as a main analytical framework as seen in Fig. 8.

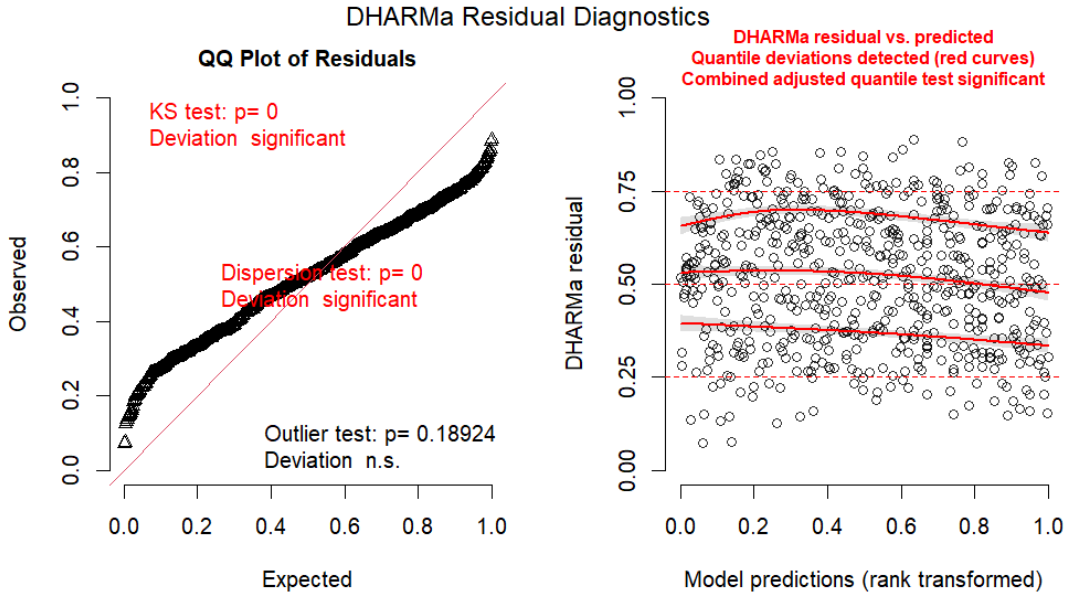


Fig. 7. Dharma Residual Diagnostics Multilevel Negative Binomial Model

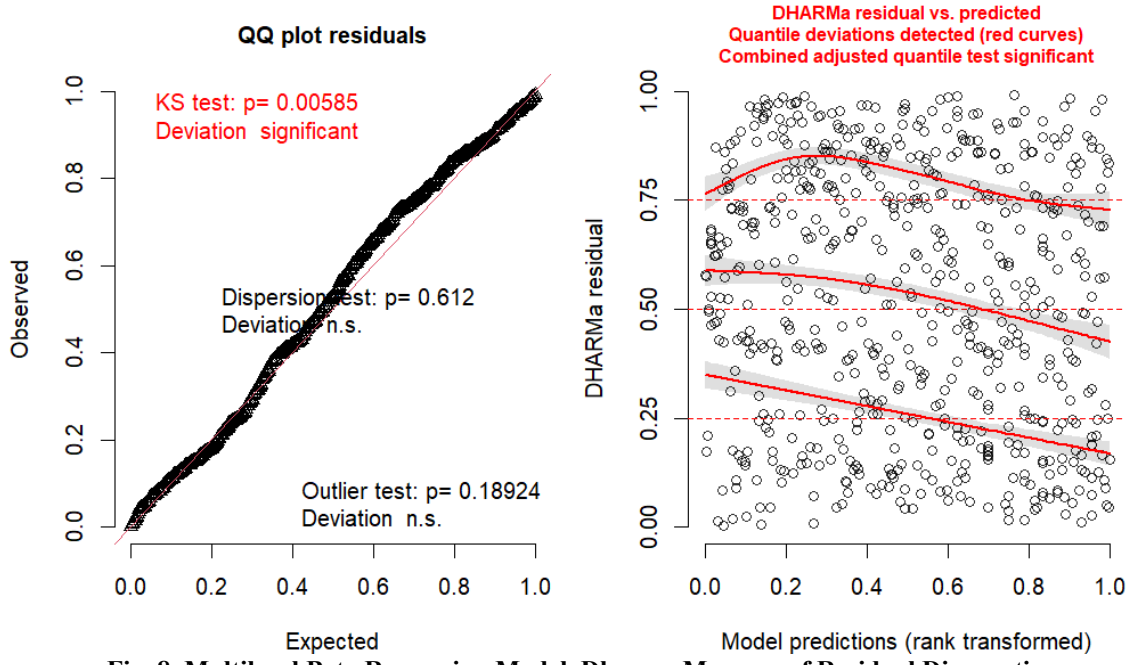


Fig. 8. Multilevel Beta Regression Model, Dharma Measure of Residual Diagnostics

Fig. 8 shows Multilevel Beta Regression model, $DHARMa$ measure of residual diagnostics. Left panel: QQ plot, there is a strong adherence to the diagonal reference line with slight upper-tail deviations ($KS\ test\ p =$

0.006; dispersion test $p = 0.612$, not significant.; outlier test $p = 0.189$, not significant). Right panel shows Residuals versus model predictions with acceptable scattered results with no systematic curvature as shown in the NB model. These diagnostics establish that the proportional loss outcome has been specified correctly in terms of its beta distribution. Unlike the NB model, the multilevel Beta regression demonstrate superior goodness of fit compared to alternative models. Since the non-significant dispersion test ($p = 0.612$) and outlier test ($p = 0.189$) indicate that the beta distribution is specified correctly in the proportional loss outcome, there is no systematic over or underdispersion. Despite being statistically significant, the KS test is statistically significant in large samples and does not imply substantive model misspecification. Adequate model fit is verified by other diagnostics (dispersion, residual plots). The plot of the residual's vs fitted values indicates that there is a satisfactory values distribution without the steep turn as noticed in the NB model. This combination of diagnostic results to a great extent confirms the hypothesis that the multilevel Beta regression is the appropriate and well-fitted model to use with this data, and all of the inferences made based on the parameter estimates are valid statistically. The single-level $betareg$ model provided a $pseudo\ R^2 = 0.7612$, which implies that the fixed effects explain the proportional losses variation of the 622 respondents in the analytic sample.

5 Conclusion and Recommendations

This study establishes that storage type, storage duration, market access conditions, transportation infrastructure, chemical use, and extension training are critical determinants of post-harvest loss along Kenya's potato farm-market value chain. Using a multilevel beta regression (MLBR) framework which is appropriate for bounded, nested, proportional loss data, we quantify these factors and derive clear policy implications. First, public and donor investment should prioritize accessible and affordable storage facilities such as dark rooms through farmer' groups and societies, and solar-powered cold storage at aggregation points along major transportation corridors. Second, post-harvest extension services require restructuring through an integrated co-delivery model that bundles training on approved sprout suppressants with simultaneous access via licensed agricultural actors. This would close the current gap between extension coverage (51.3%) and actual adoption (30.7%). Third, upgrading transportation infrastructure and introducing low-cost protective packaging for motorcycle transport is essential to reduce physical damage and distance-related losses.

Due to the cross-sectional design and regional sampling frame, causal inference and national generalizability are constrained. Future research should employ longitudinal or experimental designs, nationally representative geographic coverage, and objective post-harvest loss measurements. The MLBR method also offers promise for comparative analysis across other perishable crop value chains in sub-Saharan Africa.

Without coordinated action on these three pillars ie. storage, extension reform, and transport, technological interventions alone are unlikely to reduce post-harvest losses at scale in Kenya's smallholder potato sector.

Data Availability

The dataset utilized in this research can be provided by the author upon reasonable request.

Consent

As per international standards or university standards, Participants' written consent has been collected and preserved by the author(s).

Ethical Approval

This study was conducted with strict adherence to ethical considerations to protect the participants' rights and confidentiality. While conducting the study, all participants' personal information remained anonymous. Besides, the study ensured to take into consideration the cultural values and norms in the community. A permit to conduct the research was acquired from National Commission for Science, Technology and Innovation (NACOSTI) that allowed the gathering of data from farmers and retailers in Nakuru and Nyandarua counties.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Acknowledgement

The author would like to thank the department of Mathematics and Actuarial Science of Murang'a University of Technology for their support and guidance during the entire research process. I would like to extend my special gratitude to Dr. Ayubu Anapapa and Dr. John Mutuguta who guided and mentored me greatly during the study. Also, special thanks to my lovely wife and daughter Flossy and Natasha for their moral support.

Competing Interests

Authors have declared that no competing interests exist.

References

Abonazel, M. R., Algamal, Z. Y., Awwad, F. A., & Taha, I. M. (2022). A New Two-Parameter Estimator for Beta Regression Model: Method, Simulation, and Application. *Frontiers in Applied Mathematics and Statistics*, *7*. https://doi.org/10.3389/fams.2021.780322

Almasi, A., Eshraghian, M. R., Moghimbeigi, A., Rahimi, A., Mohammad, K., & Fallahigilan, S. (2016). Multilevel zero-inflated Generalized Poisson regression modeling for dispersed correlated count data. *Statistical Methodology*, *30*, 1–14.

Ariong, R. M., Okello, D. M., Otim, M. H., & Paparu, P. (2023). The cost of inadequate postharvest management of pulse grain: Farmer losses due to handling and storage practices in Uganda. *Agriculture and Food Security*, *12*(1). https://doi.org/10.1186/s40066-023-00423-7

Chen, V. Y. J., & Yang, Y. C. (2024). Geographically weighted regression analysis for nonnegative continuous outcomes: An application to Taiwan dengue data. *PLoS ONE*, *19*(12 December). https://doi.org/10.1371/journal.pone.0315327

Devine, S., Uanhoro, J. O., Otto, A. R., & Flake, J. K. (2024). Approaches for Quantifying the ICC in Multilevel Logistic Models: A Didactic Demonstration. *Collabra: Psychology*, *10*(1), 3175–3199. https://doi.org/10.1525/collabra.94263

FAO, (2019). The State of Food and Agriculture 2019. https://www.fao.org/family-farming/detail/en/c/1245425/

Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Geissinger, E. A., Khoo, C. L. L., Richmond, I. C., Faulkner, S. J. M., & Schneider, D. C. (2022). A case for beta regression in the natural sciences. *Ecosphere*, *13*(2). https://doi.org/10.1002/ecs2.3940

Gill, J. S., & Sharma, S. (2021). Post-harvest losses of cereals in developing countries: a review. *Canadian Journal of Agricultural and Applied Sciences*, *1*(1), 01-08. https://www.cannagri.ca/assets/pdf/volume-cjaas/Post-Harvest-Losses-of-Cereals-in-Developing-Countries-A-Review.pdf

Khan, Md. K., & Hasan, Md. T. (2022). A Poisson regression approach to modeling traffic accident frequency in urban areas. *American Journal of Interdisciplinary Studies*, *03*(04), 117–156. https://doi.org/10.63125/wqh7pd07

Levaj, B., Pelaić, Z., Galić, K., Kurek, M., Ščetar, M., Poljak, M., Dite Hunjek, D., Pedisić, S., Balbino, S., Čošić, Z., Dujmić, F., & Repajić, M. (2023). Maintaining the Quality and Safety of Fresh-Cut Potatoes

(*Solanum tuberosum*): Overview of Recent Findings and Approaches. In *Agronomy* (Vol. 13, Number 8). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/agronomy13082002

Ma, Y., Sanchez, V., & Guha, T. (2025). ZIP: Scalable Crowd Counting via Zero-Inflated Poisson Modeling. http://arxiv.org/abs/2506.19955

Marina Pare, B. A. (2025). Farmers' groups and the intensity of cereal-based agroforestry practice in Burkina Faso. *Forests Trees and Livelihoods*, *34*(4), 301–314. https://doi.org/10.1080/14728028.2025.2526344

Mburu, J. M., Oyugi, M. A., & Makenzi, P. M. (2023). Influence of Access to Extension Services on Smallholder Farmers' Use of Certified Seed Potatoes (CSPs) in Kipipiri Sub-County, Kenya. *Asian Journal of Agricultural Extension, Economics & Sociology*, *41*(4), 169–177. https://doi.org/10.9734/ajaees/2023/v41i41891

Meheretu, Y., Ashenafi, F., Kidane, D., & Welegerima, K. (2022). Stored-grain losses and management practices among smallholder cereal farmers in Northern Ethiopia. *International Journal of Pest Management, 68*(1), 59–68. https://doi.org/10.1080/09670874.2020.1795301

Morris, D. S., & Sellers, K. F. (2022). A Flexible Mixed Model for Clustered Count Data. *Stats*, *5*(1), 52–69. https://doi.org/10.3390/stats5010004

Mugo, J. N., Karanja, N. N., Gachene, C. K., Dittert, K., Gitari, H. I., & Schulte-Geldermann, E. (2021). Response of potato crop to selected nutrients in central and eastern highlands of Kenya. *Cogent Food and Agriculture*, *7*(1). https://doi.org/10.1080/23311932.2021.1898762

Mwakidoshi, E. R. (2021). Economic Importance, Ecological Requirements and Production Constraints of Potato (*Solanum tuberosum* L.) in Kenya. *International Journal of Bioresource Science*, *8*(2). https://doi.org/10.30954/2347-9655.02.2021.1

Olakiumide, O. (2021). Post-Harvest Loss Reduction: Enhancing Food Security and Economic Sustainability. https://mail.hdpublication.com/index.php/jss/article/view/169

Saha, A., Gandrakota, A., & Morozov, A. V. (2026). Poisson Log-Normal Process for Count Data Prediction. http://arxiv.org/abs/2602.05926

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological methods, 11(1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Zsögön, A., Peres, L. E. P., Xiao, Y., Yan, J., & Fernie, A. R. (2022). Enhancing crop diversity for food security in the face of climate uncertainty. *Plant Journal*, *109*(2), 402–414. https://doi.org/10.1111/tpj.15626

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). This publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© Copyright (2026): Author(s). The licensee is the journal publisher. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
https://pr.sdiarticle5.com/review-history/156473