Kenya Evaluation

José I. Velarde Morales

February 15, 2023

1 Baseline Method

We evaluate our method by comparing its performance with the method developed by Chantarat et al. (2013). This method is the standard method used in academic publications describing the design of index insurance contracts (see Flatnes, Carter, and Mercovich (2018); Jensen et al. (2019)). It is also what was used to design Kenya's Index Based Livestock Insurance (IBLI) program. In what follows, we will be calling the smallest unit of observation a "location." In practice, this could be a village or a group of villages. When designing index insurance, these smaller units of observation are assigned to larger groups, or zones.

1.1 Method description

- First, locations are assigned to zones. Assignments are either taken as exogenously given, or a clustering algorithm is used to group locations into zones. This clustering is usually based on historical weather data.
- Historical data is then used to fit a linear regression model to predict losses in each cluster. A different model is estimated for each cluster.
- Contracts are of the form: $I(\theta) = \max(\hat{\ell}(\theta) - \ell^*, 0) \times TIU \times P_{IU}$ where $\hat{\ell}(\theta)$ is the predicted loss rate, ℓ^* is the strike value, TIU is the total number of insured agricultural units, and P_{AU} is the price per insured agricultural unit. In other words, their contract pays farmers for the full predicted loss beyond a threshold, ℓ^*. This threshold, ℓ^* is the contract's strike value.
- The next step is to set the strike value. The method chooses the strike value that would explain the highest share of insured losses in the historical data. Specifically, the method runs the following regression: $y_s = \beta_s \hat{y_s} + \epsilon$ where y_s is the actual insured losses at strike value s and $\hat{y_s}$ is the predicted insured losses at strike value s. For example, suppose that $TIU = 100$ (ie there are 100 insured units), and that $P_{IU} = 25$ (ie each unit is worth 25), and that $\ell^* = 0.25$ (ie contract starts paying out once the predicted loss rate exceeds 25%). If the actual loss rate is 0.5, then actual insured losses would be $y_{25} = \max(\ell - \ell^*, 0) \times TIU \times P_{IU} = (0.5 - 0.25) \times (100) \times (25)$. If the predicted mortality rate in that scenario was 0.4, the predicted insured losses, $\hat{y_{25}} = \max(\hat{\ell}(\theta) - \ell^*, 0) \times TIU \times P_{IU} = (0.4 - 0.25) \times (100) \times (25)$. The method uses historical data to calculate $y_s, \hat{y_s}$, and then runs the following regression: $y_s = \beta_s \hat{y_s} + \epsilon$. The method chooses the strike value $s = \arg\max_s \beta_s$. The goal of choosing the strike value that explains the largest share of the losses is to minimize the basis risk, which is the probability that a loss occurs but that the insurance contract doesn't pay out. This takes into account the fact that the prediction model, $\hat{\ell}(\theta)$ might be better at predicting some

losses better than others. For example, we could have a prediction model that is good at predicting large losses, but bad at predicting small losses.

2 Data Sources

2.1 Kenya Household Survey Data

This survey was conducted as part of the development of the Index Based Livestock Insurance (IBLI) program in Northern Kenya. This dataset has information on household location, livestock levels and changes for each month in 2010-2013. There are 900 households in this dataset.

2.2 NDVI Data

The Normalized Difference Vegetation Index (NDVI) is a satellite-based indicator for the amount and health of vegetation. We use NDVI data for Kenya between 2000-2015.

3 Evaluation Datasets

3.1 Prediction Model Data

To create the data to train the prediction model, we first calculated the village level mortality rates for each season in our dataset. The seasons are

3.2 Augmented Household Data

4 Evaluation

4.1 Model Parameters

4.2 Evaluation Procedure

4.3 Performance Metrics

5 Model

We are interested in evaluating the following model for desigining agricultural index insurance:

Model Parameters

- ϵ: This is the ϵ used for the $CVaR$ objective.
- ϵ_K: This is the epsilon used in the formula for required capital. Recall that the required capital $K(I(\theta)) = CVaR_{1-\epsilon_K}(I(\theta)) - E[I(\theta)]$.
- c_κ: cost of capital.
- s_z: total insured amount for zone z.

Model In the model below, our objective is the maximum conditional value at risk of the net loss across all zones. The second constraint is the formula for the premium. The formula for required capital was also changed to include the sum of payouts across all zones. $\ell, \pi, I(\theta)$ are in terms of rates. So, for example, ℓ represents the share of the agricultural product that was lost. π is also expressed as a share of the total insured amount. And similarly, $I(\theta)$ is the share of the insured amount that is paid out.

$$\min_{a,b,K,\pi} \max_{z} \quad CVaR_{1-\epsilon}\left(s_z\left(\ell_z + \pi_z - \underline{I_z(\theta_z)}\right)\right) \tag{1}$$

$$\text{s.t. } \pi_z = \mathbb{E}\left[\overline{I_z(\theta_z)}\right] + \frac{1}{\sum_z s_z} c_\kappa K \tag{2}$$

$$K = CVaR_{1-\epsilon_K}\left(\sum_z s_z \overline{I_z(\theta_z)}\right) - \mathbb{E}\left[\sum_z s_z \underline{I_z(\theta_z)}\right] \tag{3}$$

$$\overline{I_z(\theta_z)} = \max\left\{0, a_z\hat{\ell}_z(\theta_z) + b_z\right\} \tag{4}$$

$$\underline{I_z(\theta_z)} = \min\left\{a_z\hat{\ell}_z(\theta_z) + b_z, 1\right\} \tag{5}$$

$$\pi_z \leq \overline{\pi_z} \tag{6}$$

6 Evaluation

6.1 Main Changes

changed the data generating process in the simulations so that ℓ can be a rate in our simulation, instead of an absolute value. One consequence of this change is that our simulation will no longer test the case where the underlying prediction model is corectly specified. However, I expect the linear model to be a good approximation in some of the scenarios we test. We will not be testing the case where the prediction model is correctly specified because we will be using the same prediction model as the baseline method, and their prediction model is a linear model. If we wanted to test the case where the prediction model is correctly specified, we woudl have to draw data from a linear model. However, if we draw our simulated data from a linear model, the outcome variable is not guaranteed to be between 0 and 1, which wouldn't make sense for the simulation.

6.2 Simulation Set Up

In this section, we describe how we set up the simulation used to evaluate our method. We describe the data generating process, the scenarios we test, and the simulation itself. We test our model on a toy example consisting of two zones. The goal of this exercise is to provide a fair comparison of the two methods. As a result, we will make sure that the two methods have the same budget constraint.

6.3 Old Data Generating Process

For the two zone example, we generate samples from two models: a linear model and a non-linear model. We choose a quadratic model for the non-linear model because it is the simplest non-linear model. In this DGP, ℓ is the value of the loss. So, if what we are simulating is livestock loss, $\ell = 5$ would correspond to 5 livestock lost. Since our prediction model is a linear model, this will allow us to evaluate how the two models perform in cases where the prediction model is misspecified. The data generating processes are as follows:

- Linear DGP: $\ell = \beta\theta + \epsilon$
- Nonlinear DGP: $\ell = \beta\theta^2 + \epsilon$

In both cases we have: $\theta \sim \mathcal{N}((5,5), \Sigma), \beta = diag(1.5, 1.5), \epsilon \sim \mathcal{N}(0, I)$.

6.3.1 New Data Generating Process

We will assume we have two equal sized zones, and for simplicity we will set $s_z = 1, \forall z$. In this model, ℓ_z is the covariate loss for zone z, expressed as a rate. It represents the loss rate for zone z. So, if what we are simulating is livestock loss, $\ell = 0.5$ would correspond to half of all livestock being lost. We want $\ell_z \in [0, 1]$, and we want θ_z to be predictive of ℓ_z, so we simulate ℓ_z using a logisitc regression model. As before, we want to evaluate how the performance of our model depends on the quality of the underlying prediction model. We use the following data generating processes (DGPs).

- Main DGP: $\ell_z = \frac{1}{1+e^{f(\theta_z)}}$
- Linear case: $f(\theta_z) = \beta\theta + \epsilon$
- Nonlinear case: $f(\theta) = \beta_0 + \beta_1\theta + \beta_2\theta^2 + ... + \beta_n\theta^n + \epsilon$

In both cases we have: $\theta \sim \mathcal{N}((0,0), \Sigma), \epsilon \sim \mathcal{N}(0, \rho I)$, with ρ chosen to keep the signal to noise ratio constant. In both cases, β is drawn randomly. In the linear case, the linear prediction model will be a good approximation, since the logistic is approximately linear except for its tails. The nonlinear case will allow us to test the performance when the underlying prediction model has low quality predictions.

6.3.2 Optimization Model Parameters

We use the following parameters for our optimization model in the simulations:

- $\epsilon = 0.2$ We picked this because it focuses on minimizing the CVaR of the 80^{th} percentile of the loss distribution, which roughly corresponds to once in every 5 year events, which is the desired frequency of insurance payouts.
- $\epsilon_P = 0.01$ This is a commonly set value by regulators.
- $c_k = 0.15$ This is an estimate from the literature (Kielholz (2000)).
- $s_z = 1, \forall z$ This is for simplicity

6.3.3 Scenarios to be tested

We are interested in how the two models behave in three basic scenarios. The first scenario is when there is no correlation between the losses in the two insured zones. The second scenario is when the losses in the two zones are positively correlated. The last scenario is when the losses in the two zones are negatively correlated. We test both DGPs for each scenario.

No correlation Case This is the baseline case where the losses of the two zones are uncorrelated.

- $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

Positive correlation Case This is the case where the losses of the two zones are positively correlated. This is an important scenario to test, because it is highly common in agricultural insurance. Covariate risk is one of the reasons that agricultural insurance is so difficult, since it increases the risk of catastrophic losses for the insurer. Intuitively, if one farmer was affected by a drought, it is likely that many others were affected as well. This can happen when two insured areas are geographically close to each other.

- $\Sigma = \begin{bmatrix} 2 & 1.6 \\ 1.6 & 2 \end{bmatrix}$

Negative correlation Case This is the case where the losses of the two zones are negatively correlated. This is a common feature of large scale climate processes. For example, certain El Nino-Southern Oscillation states are associated with increased rainfall in the Greater Horn of Africa, but with increased drought probabiliy in Southern Africa (Barrett et al. (2007)).

- $\Sigma = \begin{bmatrix} 2 & -1.6 \\ -1.6 & 2 \end{bmatrix}$

6.3.4 Simulation details

For each scenario we draw 300 samples for training, 50 samples for parameter selection (which are used to select the strike value in the baseline model), and 100 samples for evaluation. We run 1000 of these scenarios and compute the metrics for each one. We then report the median, 5^{th}, and 95^{th} percentile values of each performance metric across the 1000 simulations.

1. Draw samples from model, samples will be of the form (ℓ, θ) where ℓ is loss and θ is the predictor.
2. Train linear prediction model. We run $\ell = \beta_0 + \beta_1\theta + \epsilon$. Use model to generate predictions $\hat{\ell}(\theta)$ for training and test data.
3. Determine the parameters for baseline contracts using method described in section 2.
4. Once the baseline contracts have been determined, use training data to determine cost of baseline method on the training data. This gives us $\overline{\pi}$ for our model.
5. Use $\hat{\ell}$ from step 2, training data, and $\overline{\pi}$ from step 4 as input into optimization model. Use optimization model to determine contract parameters.
6. Given test data, generate predictions and use predictions to calculate payouts from baseline and from optimal contract.
7. Calculate performance metrics on test data.

6.4 Performance Metrics

The following are the metrics we calculate on the test set. Below, N is the sample size.

Maximum CVaR This is the maximum Conditional Value at Risk of the farmer's net loss across the two insured zones. For each sample $\left\{\ell_1^i, \theta_1^i, \ell_2^i, \theta_2^i\right\}_{i=1}^N$ in the test set, we calculate the net loss, $\Delta\ell_j^i \triangleq \ell_j^i + \pi_j - I_j(\theta_j^i)$. We then compute the $(1-\epsilon)-$quantile of this quantity by zone, and then calculate the average of all values greater than or equal to this quantity by zone. We then take the maximum $\max\left\{CVaR_{1-\epsilon}\left(\Delta\ell_1\right), CVaR_{1-\epsilon}\left(\Delta\ell_2\right)\right\}$.

Maximum VaR This is the maximum Value at Risk of the farmer's net loss across the two insured zones. For each sample $\left\{\ell_1^i, \theta_1^i, \ell_2^i, \theta_2^i\right\}_{i=1}^N$ in the test set, we calculate the net loss, $\Delta\ell_j^i \triangleq \ell_j^i + \pi_j - I_j(\theta_j^i)$. We then compute the $(1-\epsilon)-$quantile of this quantity by zone, and take the maximum $\max\left\{VaR_{1-\epsilon}\left(\Delta\ell_1\right), VaR_{1-\epsilon}\left(\Delta\ell_2\right)\right\}$.

Difference in VaR This is the difference in Value at Risk of the farmer's net loss in the two insured zones. For each sample $\left\{\ell_1^i, \theta_1^i, \ell_2^i, \theta_2^i\right\}_{i=1}^N$ in the test set, we calculate the net loss, $\Delta\ell_j^i \triangleq \ell_j^i + \pi_j - I_j(\theta_j^i)$. We then compute the quantile of this quantity by zone, and take the absolute value of the difference: $\left|VaR_{1-\epsilon}\left(\Delta\ell_1^i\right) - VaR_{1-\epsilon}\left(\Delta\ell_2^i\right)\right|$.

Maximum Semi-Variance

Maximum Skewness

Required Capital We report this measure because we think it is one of the comparitive advantages of our method, and it has implications for the insurer. Higher capital requirements for the insurance translate to higher costs for the insurer, and it is cost that is not necessarily benefitting the farmers. The formula for required capital is: $K(I(\theta)) = CVaR_{1-\epsilon_P}(\sum_z s_z I_z(\theta)) - \mathbb{E}[\sum_z s_z I_z(\theta)]$. For the $CVaR_{1-\epsilon_P}$, we first calculate the sum of all payouts in every scenario. We use these sums to calculate the empirical $VaR_{1-\epsilon_P}(\sum_z s_z I_z(\theta))$. We then calculate the average of all sums greater than or equal to this quantity. We calculate $\mathbb{E}[\sum_z s_z I_z(\theta)]$ using the empirical mean. We set $\epsilon_P = 0.01$ because it is a commonly used value by regulators.

Average Cost of Insurance We report this measure to ensure that the two methods have the same (or very similar costs). This will make it easier to compare the methods. We define this to be $\frac{1}{N}\sum_{i=1}^N \sum_z I_z(\theta_z^i) + c_\kappa K$. This is the empirical average of the cost of the insurance in every scenario in the test set plus the cost of capital

Probability of loss exceeding threshold We report this metric because it is of particular interest to practitioners. This metric is motivated by the literature on poverty traps, which shows that negative shocks can have very long lasting effects for individuals, especially if these shocks bring the individual's wealth below a certain threshold. We calculate this in the following way. For each sample $\left\{\ell_1^i, \theta_1^i, \ell_2^i, \theta_2^i\right\}_{i=1}^N$ in the test set, we calculate the net loss, $\Delta\ell_j^i \triangleq \ell_j^i - I_j(\theta_j^i)$. We then create an indicator variable $p_j^i = \mathbb{1}\{\Delta\ell_j^i > \bar{\ell}\}$. The probability of loss exceeding a certain threshold is then: $P(\Delta\ell > \bar{\ell}) = \frac{1}{N}\sum_{i=1}^N (p_1^i \vee p_2^i)$. We set $\bar{\ell}$ to be the 60^{th} percentile of the loss variable, ℓ.