What I've been stuck on

José I. Velarde Morales

November 28, 2022

1 Overview

In this document, I describe the state of our evaluation, and what I've been stuck on this last week or so. I first describe the old model and evaluation approach, along with its assumptions and shortcomings. Next, I describe the updated model. I then describe what we want out of the evaluation. I also describe an approach widely used in actuarial sciences, and the difficulties I've had adapting it to our setting. I conclude with my proposed approach to evaluating the new model.

2 Old Approach

2.1 Loss Value Model

The model below is our previous model. In this model, ℓ_z represents the loss in zone z, P_z is the maximum payout amount in zone z, and B is the budget constraint for the expected cost of the insurance.

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

$$\text{s.t. } \mathbb{E}\left[\sum_z \overline{I_z}(\theta_z)\right] + c_\kappa K \leq B \tag{2}$$

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

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

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

2.2 Evaluation Approach

We evaluated the model using the following two zone example. 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. 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: $\theta \sim \mathcal{N}((5,5), \Sigma), \beta = diag(1.5, 1.5), \epsilon \sim \mathcal{N}(0, I)$.

2.3 Shortcomings

- The model didn't incorporate the price the farmers paid for the insurance.
- It was difficult to include the premium in the model, because the premium is generally expressed as $\mathbb{E}[I_z(\theta_z)] + c_k K$, but the expected payout is taken across farmers of different sizes in a zone. This would mean that all farmers in the same zone would pay the same price, regardless of size.

3 Updated Model

3.1 Motivation

We updated the model to use rates instead of absolute values. This allowed us to more easily incorporate the premium into the objective, and it made it easier to incorporate cases where there are farmers of radically different sizes in a single zone. It also allowed us to express the premium as a share of the total insured amount, which is how premiums are expressed in practice.

3.2 Loss Rate Model

This is the updated model. In this model, $\ell, \pi, I(\theta)$ are all rates. So, for example, ℓ represents the share of the product that was lost. π is the price the farmer pays for the insurance, as a share of the total insured amount. We can think of π_z as being the price per insured unit. And similarly, $I(\theta)$ is the share of the insured amount that is paid out. s_z is the amount insured in zone z. The first constraint is the definition of the premium, the second constraint is the definition of the required capital, and the last two constraints are approximations we used for $I_z(\theta_z)$ to make the problem convex.

$$\min_{a,b,K,\pi} \max_{z} \quad CVaR_{1-\epsilon}\left(s_z\left(\ell_z + \pi_z - \underline{I_z(\theta_z)}\right)\right) \tag{6}$$

$$\text{s.t. } \pi_z = \mathbb{E}\left[\overline{I_z(\theta_z)}\right] + \frac{1}{\sum_z s_z} c_\kappa K \tag{7}$$

$$K = CVaR_{1-\epsilon_K}\left(\sum_z s_z \overline{I_z(\theta_z)}\right) - \mathbb{E}\left[\sum_z s_z \underline{I_z(\theta_z)}\right] \tag{8}$$

$$\overline{I_z(\theta_z)} = \max\left\{0, a_z \hat{\ell}_z(\theta_z) + b_z\right\} \tag{9}$$

$$\underline{I_z(\theta_z)} = \min\left\{a_z \hat{\ell}_z(\theta_z) + b_z, 1\right\} \tag{10}$$

$$\pi_z \leq \overline{\pi_z} \tag{11}$$

3.3 What we want in an evaluation

- We want to simulate the loss rates for the two zones with varying degrees of correlation. More concretely, we want to simulate loss rates that are independent, positively correlated, and negatively correlated.
- We want θ_z to be predictive of the loss rate, ℓ_z. This is because unlike in traditional insurance where ℓ_z is observed, in index insurance we use a signal, θ_z to create a prediction of the loss, $\hat{\ell_z}$.

4 Actuarial Evaluation Approach

4.1 Copulas

A copula is a multivariate distribution that has uniform marginal distributions. Copulas allow you to sample from a multivariate distribution where you specify the correlation between the variables, but the variables are allowed to have different marginal distributions. So, for edxample, it would allow you to smaple from the joint distribution of two variables, X_1, X_2, with correlation $\rho = 0.7$ and where X_1 follows a Gumbel distribution and X_2 follows a Pareto distribution.
Copulas are widely used in actuarial science and quantitative risk management. In actuarial science, the general method is to fit an appropriate distribution to each zone's loss severity, and then use a copula to draw from the joint distribution. So, for example, if an analyst wants to simulate the loss severity in two zones, they might decide that the losses of the first zone follow a Pareto distribution, and that the losses of the second zone follow a Weibull distribution. They would first fit each distribution independently, specify the correlation between the two zones, and then use a copula to draw from the joint distribution of the losses of th two zones. This joint distribution would have the desired marginal distributions (Pareto and Weibull), and the desired correlation between zones.

Popular distributions for modeling loss severity The following are some distributions that are commonly used to model loss severity in actuarial science:

- **Gamma:** $p(x) = \frac{(x/\theta)^\alpha}{x\Gamma(\alpha)} e^{-x/\theta}$
- **Pareto:** $p(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}$
- **Weibull:** $p(x) = \frac{\alpha}{\theta} \left(\frac{x}{\theta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\theta}\right)^\alpha\right)$

Benefits This approach gives you a lot of modeling flexibility, because the distribution of the losses of the individual zones can be different. It also has the advantage of being widely used in industry.

Shortcomings I have been having a hard time incorporating this into our evaluation because there are some key differences between our setting and the traditional insurance setting. First, the traditional insurance setting assumes that losses are perfectly observed, so there is less of a worry of payouts not correlating well with losses. The focus of actuarial analyses is often just the risk involved in offering the insurance, in which case you just need to model the insurer's losses, which are assumed to be basically the same as the insured party's realized losses. In our setting, we make payouts based on our predictions of ℓ. As a result, our simulations need to take into account the fact that we are predicting ℓ using θ. In other words, we need θ to be predictive of ℓ in our simulations. It is not clear to me how to incorporate predictions into this approach, or to the distributions that are commonly used in this setting.

5 Proposed Approach

This is the best option that I have been able to come up with.

5.1 Logit DGP

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 logit model. As before, we want to evaluate how the performance of our model depends on the quality of the underlying prediction model. We will still use a linear prediction model, since it's what's used by the baseline method we are comparing against, and the focus of the project is the contract design step, not the prediction step. We use the following data generating processes (DGPs).

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

In both cases we have: $\theta \sim \mathcal{N}((0, 0), \Sigma), \epsilon \sim \mathcal{N}(0, \eta I)$, with η chosen to keep the signal to noise ratio constant, and Σ chosen to yield the desired correlation between losses. 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.

Benefits This approach has the two features we wanted in the evaluation. It allows us to predict ℓ using θ, it allows the loss rates to have different degrees of correlation, and it will allow us to vary the quality of the prediction model. When $f(x)$ is linear I expect the linear prediction model to perform reasonably well, and when $f(x)$ is a long polynomial I expect the performance of the linear prediction model will deteriorate. This is also guaranteed to output loss rates between 0 and 1.

Drawbacks One drawback of this approach is that we can't claim that by setting $f(x)$ to be a long polynomial we can approximate any function, which we could claim when $\ell = \beta_0 + \beta_1\theta + ... + \beta_n\theta^n + \epsilon$. However, that required that ℓ be the value of the loss, which led to other modeling issues. Another small drawback of the approach is that it doesn't use the standard actuarial distributions or methods.