SOA 2021 ILEC Mortality Prediction Contest

Jimmy Risk, Nhan Huynh, Mike Ludkovski

7/31/2021

Methodology

We worked with `R`, utilizing the `tidyR` framework which simplifies data manipulation and links with `ggplot` visualizations.

The primary aim of our approach was to combine a non-parametric correlation structure in Age and Calendar Year leveraging existing mortality models, along with a parametric dependence on the large collection of covariates available in the Individual Life Mortality Experience (ILM) dataset. To do so, we came up with a two-step procedure. In the first step, we fit an Age-specific non-parametric mortality model, linking to the external Human Mortality Database data for US mortality by gender. In the second step, we model the resulting residuals through a Generalized Linear Model to account for the multitude of covariates in the ILM.

The ILM data set contains $33,807,927$ rows of data (observations) with the following columns, generically indexed by row i below:

- D_i the number of deaths,
- E_i the number of deaths and policies exposed, and
- X_i the vector of covariates containing: `Observation_Year`, `Preferred_Indicator`, `Gender`, `Smoker_Status`, `Insurance_Plan`, `Issue_Age`, `Duration`, `Attained_Age`, `Age_Basis`, `Face_Amount_Band`, `Issue_Year`, `Number_Of_Preferred_Classes`, `Preferred_Class`, `SOA_Anticipated_Level_Term_Period`, `SOA_Guaranteed_Level_Term_Period`, `SOA_Post_level_Term_Indicator`, `Select_Ultimate_Indicator`.

The categorical covariates in X_i are encoded numerically. Furthermore, we use x_{ag} to denote `Attained_Age` and x_{yr} for `Observation_Year`.

As is typical in stochastic mortality modeling, we work with the central log mortality rate

$$m_i = \frac{D_i}{E_i}, \qquad \text{(central mortality rate)}$$
$$y_i = \log(m_i). \qquad \text{(log mortality rate)}$$

The log mortality rate y_i is considered the output for the model.

Step 1: Multi-Population Stochastic Mortality Model

In our first step we build a **stochastic mortality** model. Our aim is to provide a data-driven set of mortality projections that capture the fundamental dependence on Age and (Calendar) Year. This requires smoothing the raw observations provided in the contest dataset and producing non-parametric Age-specific mortality projections for year 2017. Recognizing the central effect induced by `Gender` and `Smoker` status, we further incorporate these covariates into the mortality term structure.

Our approach takes log-mortality as the target object to be predicted, leveraging the vast longevity modeling toolbox. At this step we aggregate across all the other covariates, and simultaneously introduce an *external*

dataset, namely the United States mortality experience from the Human Mortality Database [HMD]. The HMD provides a publicly-accessible high-quality national-level mortality data broken out for Males and Females. It does not contain any other covariates and is based on the US Census Bureau and the National Center for Health Statistics (CDC) records. Given the enormous number of rows in the ILM, our motivation is to:

- aggregate data to infer trends in the insured population, optimizing the smoothing of raw observations;
- employ the HMD population to avoid overfitting and regularize the projected Age-specific mortality;
- rely on the HMD data for 2017 to improve predictive power for the purposes of the contest.

Thus, the reason for including HMD data is two-fold. First, we borrow information from the 2017 HMD data which helps in projecting $x_{yr} = 2017$ mortality rates for the ILM data set. Second, prior studies [Bahna-Nolan 2019] document that insured data is heterogeneous across years and therefore it is beneficial to use national population data to reduce year-over-year volatility and discontinuity. In particular, [SOA ILEC 2016] specifically mentions the use of HMD data for ILM assessment.

To leverage HMD, we employ a multi-population methodology that provides flexible information *fusion* across multiple datasets. Specifically, building upon the very recent work in [Huynh & Ludkovski 2021a], we develop a multi-output Gaussian Process (MOGP) model. GP modeling takes a spatio-temporal approach, representing log-mortality as a data-driven response surface indexed by x_{ag}, x_{yr} and x_{pop} which is a factor covariate encoding the different sub-populations being considered. The model infers a correlation structure that governs the dependence among different mortality rates based on the distance between respective cells in x_{ag} and x_{yr}, scaled by the length-scales θ_{ag}, θ_{yr}. It moreover infers a cross-correlation between populations linking respective mortality experiences. The MOGP-predicted log-mortality can be interpreted as a linear combination of all observed mortality rates, weighted by the model-driven similarity between different (Age, Cal Year) pairs. This gives a non-parametric, data-driven fit that jointly smooths the noisy observed mortality rates in `Age` and `Year`, as well as fuses information from the HMD.

Models Developed: In total, we have $2 \cdot 3 + 2 = 8$ populations, 6 populations from ILEC that are broken out by Gender (2 factor levels) and Smoker Status (3 levels), plus 2 populations (Males/Females) from HMD. In terms of span in Ages and Years we have:

– Individual Life Mortality Experience data, $0 \le x_{ag} \le 120$, $2009 \le x_{yr} \le 2016$;

– Human Mortality Database data, $0 \le x_{ag} \le 109$, $2009 \le x_{yr} \le 2017$. (In fact, HMD presently covers 1933–2019, so we could potentially include that additional information as well.)

Given the clear evidence of varying mortality experience for Males and Females, we construct separate gender-based models (in other words we assume that the experiences for Males and Females are statistically conditionally independent). At the same time, given the expected strong dependence between different Smoking statuses, we build a *joint* model across those sub-populations. Thus, we consider a joint model for {HMD Males, ILEC Smoker Males, ILEC Non-Smoker Males, ILEC Unknown Males}, encoded as $x_{pop} = 1, 2, 3, 4$ respectively, all treated via a *single* MOGP model fitted utilizing data from all stated populations.

We considered breaking additionally according to `Preferred_Indicator` or `Select_Ultimate_Indicator`; any discrete variable can technically be used to split. However, adding sub-populations requires to consider all possibilities, exponentially increasing the number of populations in the MOGP. This is computationally expensive (and makes the MOGP less stable) and likely to yield limited benefits. Thus, for tractability we stick with treating `Gender` separately and jointly handling `Smoker` (experiments fusing by `Gender` as well, which give a 8-population MOGP, produced comparable results). Our ongoing work [Huynh & Ludkovski 2021b] considers hierarchical multi-population mortality models that could be utilized if we had more time.

In order to handle the highly non-stationary correlation structure in `Age`, we build 3 different age-segmented models covering Ages 0-30, 30-70, 70-100. To avoid discontinuity around the `Age` cut-overs, we train the models on overlapping `Age` segments $[0, 35], [25, 75], [65, 100]$, and Years 2009-2016 for ILM and 2009-2017 for HMD. One feature of the MOGP model is that it straightforwardly handles such "notched" input that

has different observations for different sub-populations (in particular, no $x_{yr} = 2017$ for ILM and no $x_{ag>109}$ for HMD). In essence, it "borrows" the information from $x_{yr} = 2017$ in the HMD data set, and uses the correlations found with the ILM sub-populations to assist in predicting $x_{yr} = 2017$ for ILM.

For Ages 100+ the ILEC data is very sparse and noisy. Since those rows contribute little to overall predictive accuracy, we postponed the challenge of modeling them and assumed a constant mortality rate for all ages above 100, set to the smoothed mortality rate at $x_{ag} = 100$. See the flat sections in Figure 1 below.

In sum, we build 6 distinct MOGP models, each for 4 sub-populations (3 from ILEC, 1 from HMD). For the purposes of our subsequent exposition, we bind them back across the Age segments, and present the results in terms of `Gender` and `Smoker`.

MOGP Details: Our MOGP is implemented using `DiceKriging` and `kergp` packages in `R`. We select a constant GP prior trend and the Squared-Exponential covariance kernel; these are standard choices in the literature. Based on our extensive experience with GP modeling, the choice of this "GP mean function" makes little impact for the purposes of doing a one-year-out projection. We use our expert knowledge to restrict prior GP lengthscale ranges.

Cross-population correlations: As expected, ILEC mortality rates are substantially lower than those in the HMD. However, that gap is greatly affected by `Smoker` status. For non-smokers, our fitted model suggests over 20% reduction in mortality, while Smokers have mortality that actually exceeds that in HMD for Ages 50-85. `Unknown` status seems to be effectively the same as `Smoker` for Ages below 45 (suggesting that all non-Smokers are motivated to self-report as such) and closer to non-Smoker for Ages above 70 (perhaps due to data entry issues for older policies).

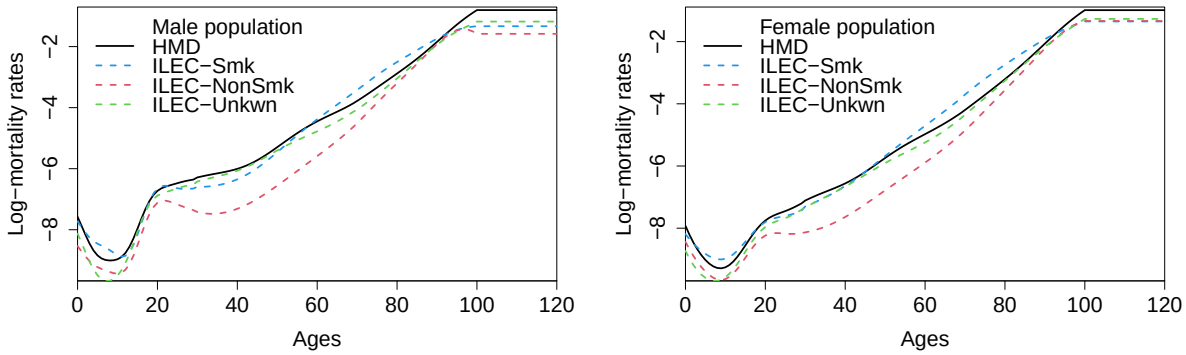

Figure 1: Projected 2017 log-mortality as a function of Age and Smoker status.

Figure 2 shows the heat maps illustrating the ratios of ILEC mortality by Smoker status compared to the HMD overall population. We see that the ratios are growing over time.

We obtained correlation of about 70% between HMD and ILEC Smoker/Non-smoker sub-populations, and about 90% for the `Unknown` sub-population. This suggests that the `Unknown` group is in fact very similar (ignoring other covariates) to overall population. We also obtained correlations of 60-90% between different ILEC sub-populations. The results are quite similar across Males/Females and across different Age segments.

Step 2: Gaussian GLM with Ridge Regression

After fitting the MOGP model, for each i (indexed by Age, Year) we obtain $\hat{y}_i^{GP}$ through the posterior mean of the Gaussian process, and we compute the residual of the log mortality rate as

$$e_i = y_i - \hat{y}_i^{GP}. \tag{1}$$

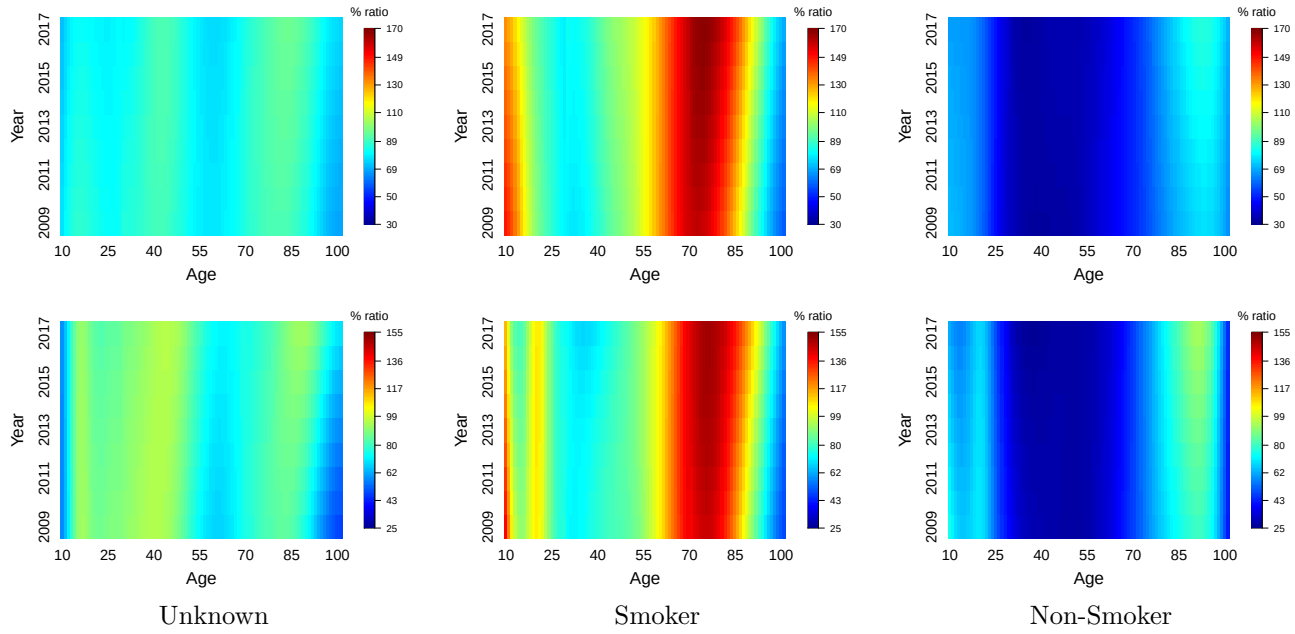

Figure 2: Ratios of ILM mortality to HMD population across Age (x-axis) and calendar Year (y-axis). *Top row*: Female; *bottom row*: Male.

Next we build a Generalized Linear Model (GLM) on the residuals e_i's. We assume that e_i is a linear function of other covariates X_i and since it is real-valued, a Gaussian GLM is appropriate. Since $m_i = \frac{D_i}{E_i}$, $e_i = \log m_i - \hat{y}_i^{GP} = \log(D_i/E_i) - \log(\hat{D}_i^{GP}/E_i) = \log(D_i/\hat{D}_i^{GP})$. In other words, $\exp(e_i) = D_i/\hat{D}_i^{GP}$, the ratio of actual to (GP) expected deaths, and the GLM is modelling the log of the A/E deaths ratio.

We did consider other alternatives that make sense from the perspective of targeting the contest metric which is not on the log-scale. We experimented with Gamma and Inverse-Gaussian GLM fitted to $\exp(e_i) = D_i/\hat{D}_i$ with respective link functions so that the log mean is linear in X_i but that estimation was very slow and did not yield better results before the deadline. In future work we would like to consider a Poisson GLM fitted to the count data $D_i \sim \text{Pois}(E_i \exp(\mu_i))$ and then fitting a GP to the residuals.

The proposed GLM assumes Gaussian errors with a linear dependence of the response on all covariates, as well as their first-order interactions:

$$e_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j=1}^{p} \sum_{k<j} \beta_{jk} x_{ij} x_{ik} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \tag{2}$$

where x_{ij} is the ith row and jth column of the data set, after appropriately coding categorical variables. This form is motivated by the large amount of categorical variables, of which various linear relationships between numerical x_{ij} and e_i may differ, for different categorical levels.

Due to the large number of resulting terms in the model, as well as many collinear variables, we choose **ridge regression** to infer the β coefficients. As we expect all covariates to have a potential effect, we did not want to exclude any covariates and hence focused on regularized regression. A shrinkage approach via LASSO is a reasonable alternative and could be more interpretable.

In estimating $\beta_j, j = 0, \ldots p$, we further opt to perform a **weighted** ridge regression with weights (w_i):

$$\hat{\beta} = \text{argmin}_\beta \left\{ \sum_{i=1}^{n} w_i \left(e_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j=1}^{p} \sum_{k<j} \beta_{jk} x_{ij} x_{ik} \right) \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{3}$$

As is common, the tuning parameter λ is obtained through 5-fold cross-validation (package `glmnet`). We carry out two iterations of (3):

Iteration 1: Perform 5-fold cross validation where in removing each fold we fit (3) with weights $w_i = E_i$, the number of exposures. The resulting model is predicted on the removed fold, obtaining a value based on the competition error metric (for each row):

$$\text{CVErr}_i = \frac{(C_i - \hat{C}_i)^2}{\hat{C}_i}. \tag{4}$$

Iteration 2: Fit a weighted ridge regression model according to Equation (3) on the full training data, using the weights $w_i = 1 + \log(1 + \text{CVErr}_i)$ based on iteration 1; those weights end up in the range of $w_i \in [1, 20]$. Based on the resulting $\hat{\beta}_j$, the GLM prediction $\hat{e}_i^{GLM}$ for any row i with covariates x_i' is

$$\hat{e}_i^{GLM} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x'_{ij} + \sum_{j=1}^{p} \sum_{k<j} \hat{\beta}_{jk} x'_{ij} x'_{ik}. \tag{5}$$

Finally, to account for observed over-estimation for rows with very low death counts, we fit an auxiliary logistic regression (with same dependence on X_i) for the event $\{D_i = 0\}$, which yields $\hat{p}_i = \mathbb{P}(D_i > 0 | X_i)$. Our final mortality projection is

$$m_i = \exp\left(\hat{y}_i^{GP} + \hat{e}_i^{GLM} \cdot \hat{p}_i\right).$$

Other regression specifications: We tested regressing on:

1. All first order terms in X_i.
2. First order terms and all first order interactions.
3. Second order terms, and all first order interactions.

Looking at the fits and residual plots, second order terms seemed unnecessary (performing worse for projecting 2016), but the interaction terms provided significant improvement over a strictly linear model.

Regression Weights: The choice of E_i as regression weight in (3) is equivalent to expanding the data into individual insureds observations rather than on the given rows. Other weighing schemes: $w_i = 1$, $w_i = \sqrt{E_i}$, $w_i = E_i^2$ were tested, but $w_i = E_i$ outperformed the others in our 2016 out-of-sample prediction.

With the choice $w_i = E_i$, we still found discernable patterns in Err_i for 2016, for example as a function of `Age`. This motivated the idea to weigh based on prediction performance and led to the 2-step iteration above. Its purpose is to prioritize minimizing predictive errors in rows where *claim predictions are underperforming*, akin to boosting. This specific weighing scheme improved 2016 out of sample prediction by approximately 10%. Other weighing schemes involving CVErr_i were also considered, but we were forced to end a full analysis due to time constraints.

One-Year-Ahead Analysis

To determine model efficacy in predicting one-year-ahead, we performed extensive analysis in fitting to 2009-2015 data and predicting on the known 2016 data. The first figure displays a metric representative of the contest

$$\text{Err} = \sqrt{\frac{\sum_{i \in \{\text{2016 data rows}\}} \frac{(C_i - \hat{C}_i)^2}{\hat{C}_i}}{\sum_{i \in \{\text{2016 data rows}\}} E_i}}, \tag{6}$$

which is computed separately over each age; smaller values are preferred. As discussed by the ILEC, the intent of this metric is to assess performance of claim amount prediction.

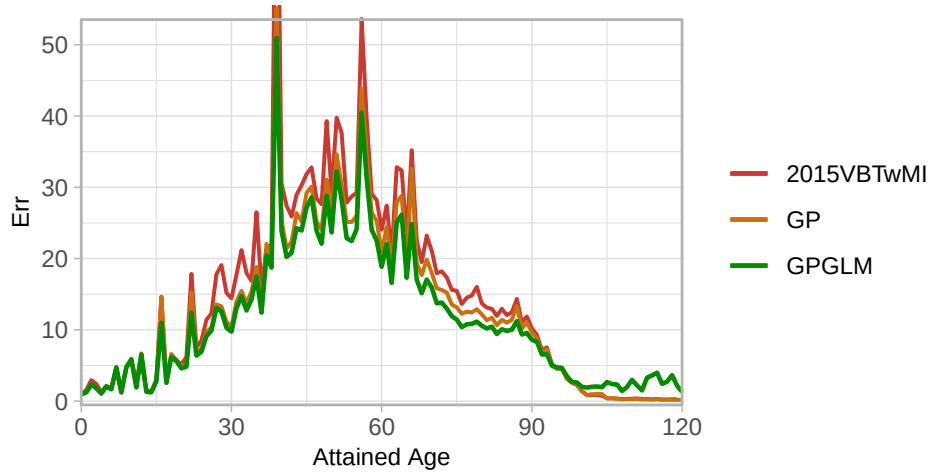

Figure 3: Cumulative Errors vs. Attained Age of Policy Holders (One-Year-Ahead 2016)

Both GP based models are shown to perform better than the 2015 VBT table. By comparing the green and yellow lines in Figure 3, the GLM portion of the model clearly provides additional error reduction. Note that the GLM can have issues in areas with sparse training data (e.g. extreme ages) since it was optimized to reduce aggregate error.

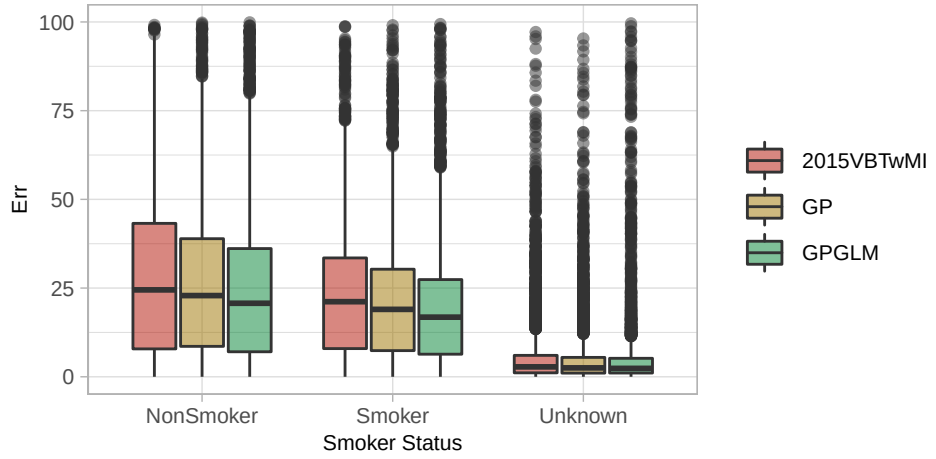

Figure 4: Cumulative Errors vs. Attained Age of Policy Holders (One-Year-Ahead 2016)

In Figure 4 the prediction errors in 2016 are first aggregated over all covariates aside from `Age`, `Duration`, and `Smoker`. A similar pattern of performance emerges as before: the GP breaks into population by smoker status, so it improves over 2015VBT. The GLM then handles additional interactions with covariates and smoker status, further improving performance over the GP.

2017 Mortality Projections

Without having policies exposed for 2017, we focus on mortality projections. The following plots display partial dependencies of predicted 2017 mortality in terms of key covariates. Within each visualization, the projections are averaged over all covariates not included and over all predicted rows to provide an average representation of the full data set given. The code used to produce the plots below will generate analogous plots for any variables by appropriately replacing `Attained_Age` and `Gender`, showcasing the efficiency of `tidyr`. Boxplots for categorical variables on the x-axis can be produced similarly.

```r
df_Plot_1 <- grid_2017_temp %>%
  group_by(Attained_Age,Gender) %>%
  summarise(mort_rate = mean(mort_rate))

ggplot(df_Plot_1, aes(x = Attained_Age, y = log(mort_rate))) +
  geom_line(aes(color = Gender)) + xlab("Attained Age") + ylab("Log Mortality Rate")
```

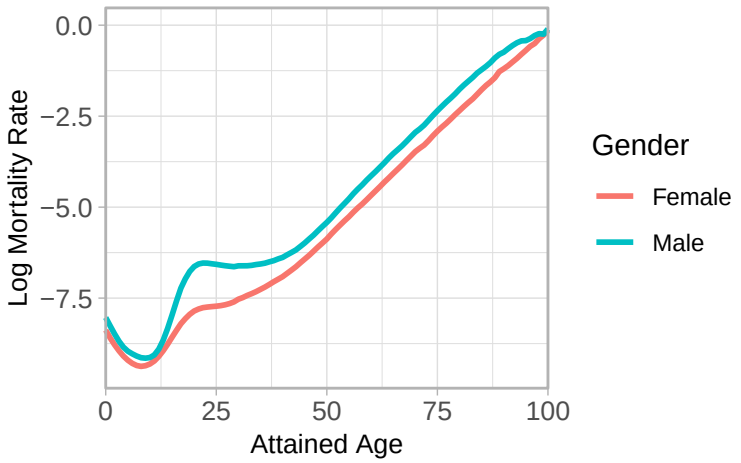

Figure 5: Partial Dependence plot of projected 2017 log-Mortality vs. Age (by Gender)

Figure 5 shows our integrated projection for 2017 log mortality. The shape is very similar to the GP fit $\hat{y}^{GP}$ shown in Figure 1. Indeed, the GLM modification should not significantly change the overall mortality shape. As expected, Male mortality is universally higher. Sparse data for extreme ages causes the two to converge around age 100.

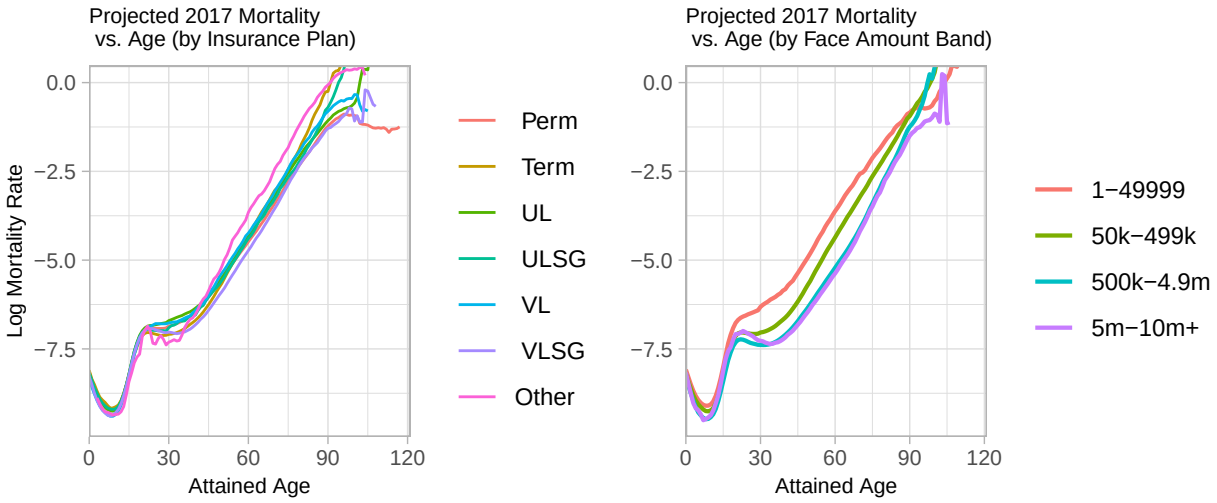

Figure 6: Partial Dependence Plots for 2017 log-Mortality in terms of Attained Age.

Figure 6 highlights the differing relationship on mortality predictions among products (left panel) and face amounts (right panel). The right panel highlights that our model detected a monotonic relationship between face band amounts: policies with higher face amounts tend to have lower mortality rates.

Figure 7 illustrates different x-axes, with issue year on the left panel and duration on the right panel. Both

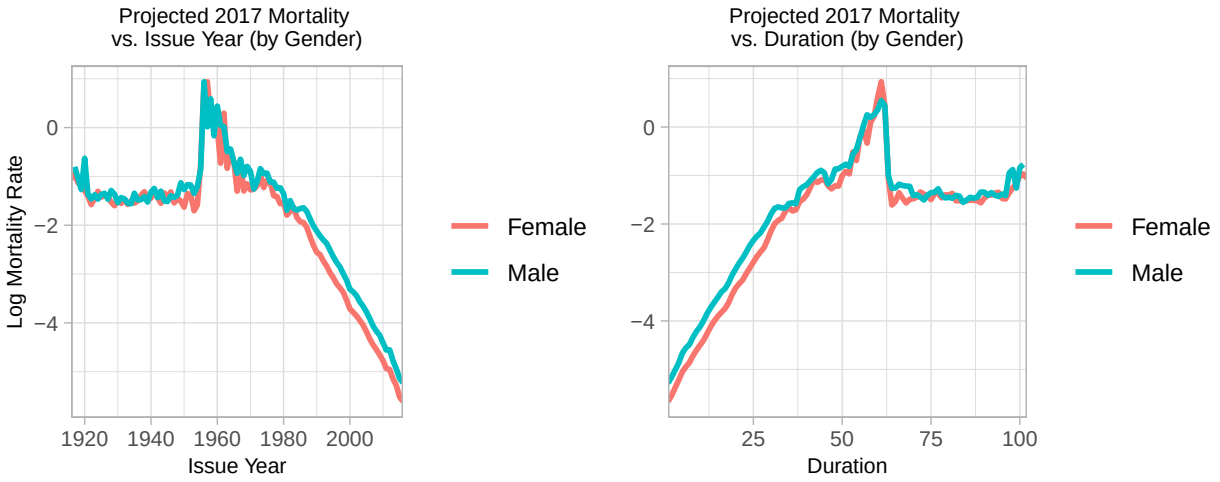

Figure 7: Partial Dependence Plots in terms of Gender.

plots have similar shapes, highlighting an inhomogeneity of the data. Furthermore, we see log mortality decreasing for higher issue years, where younger individuals tend to purchase products.

References

Bahna-Nolan, M.J. (2019). Mortality Improvement. Presentation at the SOA 2019 Life & Annuity Symposium

HMD: Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org, downloaded in June 2021. See https://www.mortality.org/hmd/USA/InputDB/USAcom.pdf for US data documentation.

Huynh, N., & Ludkovski, M. (2021a). Multi-output Gaussian processes for multi-population longevity modelling. Annals of Actuarial Science, 15(2), 318-345

Huynh, N., & Ludkovski, M. (2021b). Modeling cause-of-death data with multi-task Gaussian processes. Work in Progress.

Huynh, N., & Ludkovski, M. (2020). Tutorial on Multi-output Gaussian processes. (Includes interactive online RMarkdown notebook) https://nhanhuynh46.github.io/MOGPTutorials/

Ludkovski, M., Risk, J., & Zail, H. (2018). Gaussian Process Models for Mortality Rates and Improvement Factors. ASTIN Bulletin, 48(3), 1307-1347.

Society of Actuaries (2016). 2016 Individual Life Insurance Mortality Experience Report https://www.soa.org/globalassets/assets/files/resources/research-report/2021/2016-individual-life-report.pdf