Modeling Uncertainty in Unemployment Duration

Sneha Sarkar
School of Computing
A0304787U
e1373875@u.nus.edu

Prerana Chakraborty
School of Computing
A0305018R
e1374106@u.nus.edu

Liu Xin Ying
School of Computing
A0188304A
e0323888@u.nus.edu

Lee Jia Yun Amanda
Business School
A0296790L
amandaleejy@u.nus.edu

Akshaya Vajpeyarr
School of Computing
A0307243M
akshaya_vajpeyarr@u.nus.edu

Abstract

This paper presents a Bayesian framework for modeling the uncertainty in unemployment duration, with the goal of providing valuable insights to policymakers, job seekers, and employers. We use a combination of marginal probability distributions, iterative proportional fitting (IPF), and hierarchical modeling to estimate unemployment duration based on demographic and educational factors. The model is designed to account for uncertainty by generating credible intervals for survival predictions, thereby facilitating informed decision-making. We also implement variational inference to estimate posterior distributions and employ Monte Carlo methods for uncertainty quantification. The results demonstrate the potential of the model in understanding labor market dynamics, while also offering a flexible and scalable approach for future policy interventions. The code for this project is publicly available at this Github repository.

1 Introduction

Understanding the multifaceted nature of unemployment duration, influenced by personal, structural, and economic factors, is crucial for individual support, policy formulation, and workforce resilience. This project focuses on modeling the uncertainty surrounding unemployment duration to provide valuable insights for policymakers, job seekers (including students), and employers, enabling informed decisions and a deeper understanding of labor market dynamics. By quantifying this uncertainty, the model aims to support proactive policy interventions against long-term unemployment and can serve as a personalized career support tool suggesting re-employment strategies and promoting fairer access to retraining, while emphasizing the critical need for safeguards against discriminatory misuse, ensuring the model's insights are solely used for support and policy guidance.

2 Preprocessing

To prepare the data for modeling, we implemented a detailed preprocessing pipeline encompassing data acquisition, cleaning, marginal estimation, joint distribution modeling, and final dataset generation. The main steps are described below:

2.1 Data Sources

We utilized three official datasets from `data.gov.sg`, each contributing different marginal information relevant to unemployment statistics:

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

- **Unemployed Residents Aged 15 Years and Over by Age and Duration of Unemployment** [1]: Provides unemployment duration distributions across different age groups and sexes.
- **Unemployed Residents Aged 15 Years and Over by Highest Qualification Attained and Duration of Unemployment** [2]: Provides unemployment duration distributions across different qualification levels and sexes.
- **Median Duration of Unemployment** [3]: Provides reference median values of unemployment durations across years for calibration.

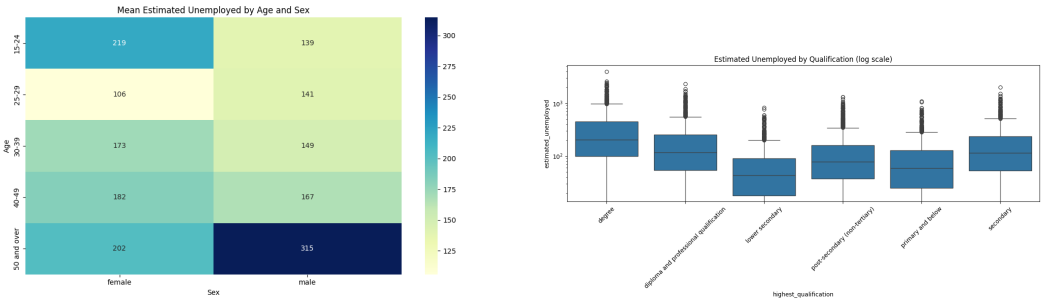
2.2 Basic Cleaning

To ensure consistency and numerical validity: Converted columns such as `unemployed` to numeric types, coercing invalid or missing entries to NaN and handling them appropriately. Mapped categorical duration ranges (e.g., "5 to 9 weeks") to corresponding numeric midpoints to allow quantitative modeling. Filled missing values carefully to avoid introducing bias in downstream calculations.

2.3 Marginal Computations

For each year, marginal tables for `age × sex` and `qualification × sex` were computed by summing the number of unemployed in each group. We derived marginal probability distributions necessary for reconstructing a joint probability table:

- Unemployed counts grouped by **year**, **age**, and **sex** from the age-duration dataset[1]
- Unemployed counts grouped by **year**, **highest qualification**, and **sex** from the qualification-duration dataset[2].



(a) Mean Estimated Unemployed by Age and Sex

(b) Estimated Number of Unemployed People by Qualification (log scale)

Figure 1: Visualizing Marginal Relationships

These visualizations helped identify missing or sparse entries in marginal data early on.

2.4 Iterative Proportional Fitting (IPF)

The IPF algorithm was used to estimate the joint distribution of age, sex, and qualification, constrained to match the observed marginals. The initial guess for the joint was a uniform array; the IPF algorithm iteratively adjusted the joint so that its marginals matched the observed ones along each dimension (age, qualification, sex). This was updated until convergence was reached or a maximum number of iterations. The result is a synthetic joint table for each year:

$$\text{Joint}_{y,a,s,q} \approx P(\text{age} = a,\ \text{sex} = s,\ \text{qualification} = q \mid \text{year} = y)$$

2.5 Duration Probability Estimation

We estimated conditional probabilities for unemployment duration through two complementary approaches. Firstly, for age-conditioned probability, for each age-sex group, we calculated the probability of each duration range: $P(\text{duration} = d \mid \text{age} = a,\ \text{sex} = s,\ \text{year} = y)$. Secondly, for qualification-conditioned probability, for each qualification-sex group, we calculated the probability

of each duration range: $P(\text{duration} = d \mid \text{qualification} = q, \text{sex} = s, \text{year} = y)$. Where marginal information was missing or sparse, probabilities were imputed carefully to ensure the model's stability.

2.6 Hybrid Probability Calculation

Recognizing that both age and qualification affect unemployment differently, we introduced a hybrid estimation approach. The hybrid probability was obtained by linearly combining the age-conditioned and qualification-conditioned probabilities with equal weighting ($w = 0.5$), reflecting no strong prior on which factor is more important: $P_{\text{hybrid}}(d) = 0.5 \times P_{\text{age}}(d) + 0.5 \times P_{\text{qual}}(d)$

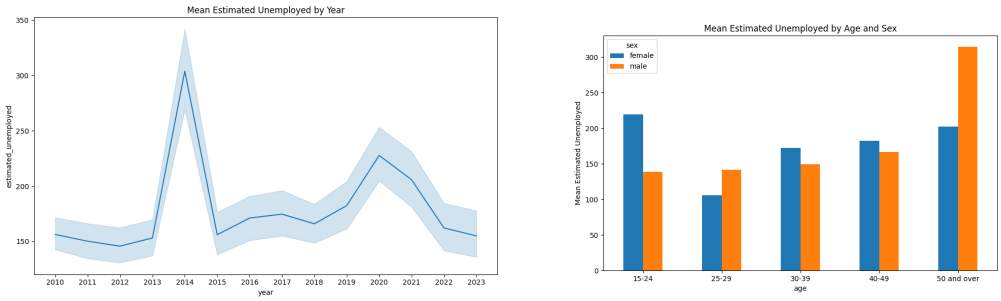
2.7 Final Estimation

After obtaining the joint distribution and hybrid probabilities, for each combination of year, age, sex, qualification, and duration, the estimated number of unemployed is:

$$\text{estimated\_unemployed} = \text{Joint}_{y,a,s,q} \times P_{\text{hybrid}}(d)$$

The resulting estimated unemployed counts were rounded to the nearest integer to produce a final dataset suitable for modeling. Finally, to validate the synthesized data, we generated:

- Boxplots showing the spread of estimated unemployed counts across qualification categories (using both linear and log scales).
- A line plot tracking mean unemployment counts across years to detect long-term temporal trends.



(a) Mean Estimated Unemployed by Year (b) Mean Estimated Unemployed by Age and Sex

Figure 2: Comparison of Mean Estimated Unemployment Trends

These visualizations confirmed the plausibility of the generated synthetic dataset before moving into model fitting.

3 Bayesian Model

The goal was to infer the distribution of unemployment duration, accounting for demographic and educational covariates, and to quantify uncertainty in predictions.

3.1 Response Variable

The midpoint of each duration interval (e.g., "10 to 14" $\rightarrow$ 12) is used as the observed duration. The response is then log-transformed: $y_i = \log(\text{duration}_i)$

3.2 Feature encoding

The input features were processed as follows:

- **Year (centered):** Each year's value was centered by subtracting the mean year across the dataset: $x_{\text{year},i} = \text{year}_i - \overline{\text{year}}$
- **Sex (binary encoded):** Encoded as 0 for male and 1 for female.
- **Age group (categorical encoded):** Each unique age group (e.g., 15–24, 25–29) was mapped to an integer index.
- **Highest qualification (categorical encoded):** Each qualification category (e.g., degree, diploma, secondary) was mapped to an integer index.

3.3 Hierarchical Linear Predictor

The expected log-duration μ_i for individual i is modeled as:

$$\mu_i = \alpha + \beta_{\text{year}}\, x_{\text{year},i} + \beta_{\text{sex}}\, x_{\text{sex},i} + \beta_{\text{age}}[x_{\text{age},i}] + \beta_{\text{qual}}[x_{\text{qual},i}]$$

where α is the intercept term; β_{year} and β_{sex} are regression coefficients for centered year and sex respectively; β_{age} is a vector of age-group-specific coefficients, indexed by $x_{\text{age},i}$; β_{qual} is a vector of qualification-group-specific coefficients, indexed by $x_{\text{qual},i}$.

3.4 Likelihood

Using a Student's t-distribution allows for greater robustness to extreme values compared to a Gaussian likelihood, which is particularly important given the presence of outliers in unemployment durations.
The observed log-duration y_i for each individual i is modeled using a Student's t-distribution, which is robust to outliers:

$$y_i \sim \text{StudentT}(\nu = 3, \mu_i, \sigma)$$

where μ_i is the expected log-duration from the hierarchical linear predictor, σ is the scale parameter, and $\nu = 3$ fixes the degrees of freedom to control the heaviness of the tails.

To account for the estimated number of unemployed individuals in each subgroup, we define weights w_i based on the estimated unemployed count:

$$w_i = \text{estimated\_unemployed}_i$$

The weighted likelihood across all individuals is given by:

$$L = \prod_i p(y_i \mid \mu_i, \sigma)^{w_i}$$

Taking the log of the likelihood (to improve numerical stability), we obtain the weighted log-likelihood:

$$\log L = \sum_i w_i \cdot \log p(y_i \mid \mu_i, \sigma)$$

3.5 Prior Distributions

We use weakly informative priors to regularize the model while allowing flexibility in capturing variability across groups. The priors for the model parameters are specified as follows:

- Intercept: $\alpha \sim \mathcal{N}(3, 2)$
- Regression coefficients for year and sex: $\beta_{\text{year}},\ \beta_{\text{sex}} \sim \mathcal{N}(0, 1)$
- Group-level effects for age: $\beta_{\text{age}} \sim \mathcal{N}(0, \sigma_{\text{age}})$ with $\sigma_{\text{age}} \sim \text{HalfNormal}(1)$
- Group-level effects for qualification: $\beta_{\text{qual}} \sim \mathcal{N}(0, \sigma_{\text{qual}})$ with $\sigma_{\text{qual}} \sim \text{HalfNormal}(1)$
- Likelihood noise scale: $\sigma \sim \text{HalfNormal}(1)$

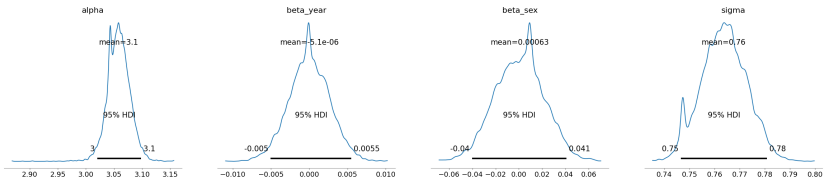


Figure 3: Posterior Parameter Plots

3.6 Model Fitting and Posterior Inference

The model is fit using the No-U-Turn Sampler (NUTS), a variant of Hamiltonian Monte Carlo, with 4 chains and 2000 posterior samples per chain (after 1000 tuning steps). The posterior samples of all parameters are saved as an ArviZ InferenceData object. The model employs interval-censored data with an adapted likelihood function to handle lower and upper bounds of unemployment duration, while incorporating log-transformations and numerical clipping to ensure stability when processing zero or negative duration values.

Interpretation of Bayesian Inference Results

The Bayesian inference results provide posterior estimates of the model parameters, including their means, standard deviations, and credible intervals. Key observations from the model are:

- The parameter `alpha` has a mean of 3.058, with a narrow credible interval (3.020, 3.098), indicating a confident estimate of this parameter.
- The effects of `year`, `sex`, and several `age` and `qualification` categories are relatively small, as their credible intervals mostly overlap zero. For instance, `beta_year` has a mean of -0.000, with a 95% credible interval ranging from -0.005 to 0.005, suggesting that the year variable does not have a strong impact on unemployment duration.
- The `sigma` parameter, representing the noise term in the model, has a relatively small standard deviation (0.009), with a credible interval between 0.747 and 0.781, indicating the model's robustness and confidence in the estimated noise level.
- The `ess_bulk` and `ess_tail` values for the parameters are generally high, confirming good convergence and mixing of the Markov chains. The `r_hat` values for all parameters are close to 1.00, suggesting that the model has successfully converged.

In summary, the model provides reliable estimates of key parameters, with some parameters (such as `year`, `sex`, and certain qualifications) showing limited effect on the unemployment duration. The overall model fit appears to be strong, with good convergence diagnostics.

3.7 Prediction for New Cases

We used the trained Bayesian model to predict unemployment durations for new cases based on posterior parameter samples. Inputs included year, sex, age group, and highest qualification. As an example, we predicted the unemployment duration for a **female aged 30–39 in 2025, with the highest qualification as "degree"**. The process involved normalizing the year, encoding categorical variables, computing a linear predictor from the posterior samples, and generating predictions by sampling from a log-normal distribution using the predictor and posterior scale parameter (sigma). The model predicts a **median** unemployment duration of **21.59 weeks**, with a **95% credible interval** ranging from **4.87 weeks to 93.66 weeks**, reflecting the inherent uncertainty in the estimation. The posterior predictive distribution for the unemployment duration is shown in the Figure 4. The plot shows the predicted unemployment duration range, capturing both the central tendency and prediction uncertainty.

The posterior predictive distribution for the unemployment duration is shown in the figure below:

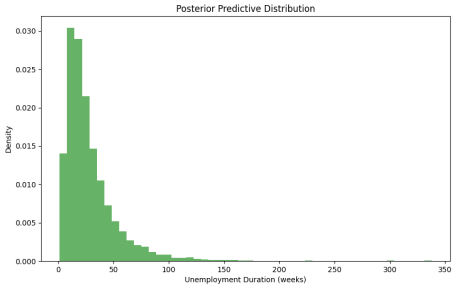


Figure 4: Posterior Predictive Distribution of Unemployment Duration for New Case

The plot illustrates the range of predicted unemployment durations for the given individual, highlighting both the central tendency and the uncertainty associated with the prediction. They provide valuable insights into the typical unemployment duration for this individual while also quantifying the uncertainty involved in the prediction process.

3.8 Visualization and Diagnostics

Bayesian model traces '.nc' and '.npy' files are on Google Drive.

- **WAIC:** Computed from 8000 posterior samples and 7560 observations log-likelihood matrix. The following values were obtained:

$$\text{elpd\_waic} = -11056.77 \quad (\text{SE} = 76.20) \qquad \text{p\_waic} = 6.39$$

- **LOO:** Computed from 8000 posterior samples and 7560 observations log-likelihood matrix:

$$\text{elpd\_loo} = -11056.77 \quad (\text{SE} = 76.20) \qquad \text{p\_loo} = 6.39$$

Additionally, the **Pareto k diagnostic values** were calculated to assess the reliability of the LOO estimates:

- **Pareto k values:**

$$\begin{aligned} (-\infty, 0.70] \quad &\text{(good):} \quad 7552 \quad (99.9\%) \\ (0.70, 1] \quad &\text{(bad):} \quad 2 \quad (0.0\%) \\ (1, \infty) \quad &\text{(very bad):} \quad 6 \quad (0.1\%) \end{aligned}$$

These results indicate that the model performs well, with a low WAIC and LOO score, and no problematic Pareto k values that would suggest issues with model fit or inference quality. Posterior Predictive checks are also done for each posterior sample. Predicted durations are generated for all groups. The predicted and observed distributions are compared using summary statistics (mean, median, std) and kernel density estimates (KDEs).This compare the observed and predicted distributions of unemployment duration.

Statistic	Observed	Predicted
Mean	27.22 weeks	28.53 weeks
Median	22.00 weeks	21.31 weeks
Standard Deviation	22.05 weeks	25.42 weeks

Table 1: Summary Statistics Comparison

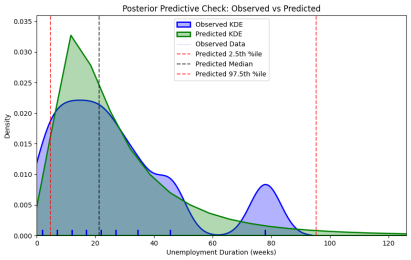


Figure 5: Posterior Predictive Check: Observed vs Predicted

4 Survival Analysis

4.1 Bayesian Modeling Approach

The goal is to infer the **posterior distribution** over weights $p(w \mid D)$, given the data $D = \{(x_i, y_i)\}_{i=1}^{N}$. According to Bayes' theorem:

$$p(w \mid D) = \frac{p(D \mid w)\, p(w)}{p(D)},$$

where:

- $p(D \mid w)$ is the likelihood of the data given weights,
- $p(w)$ is the prior over weights,
- $p(D)$ is the evidence (marginal likelihood).

In this context:

- D represents the observed dataset, where each x_i is the input feature vector (including categorical variables such as year, qualification, age, sex, and continuous variables like estimated unemployed), and y_i is the corresponding unemployment duration (survival time).
- w represents the collection of neural network weights and biases, which are treated as random variables in the Bayesian framework for **estimating the conditional distribution of unemployment duration y given the input features x, i.e.,** $p(y \mid x)$.

4.2 Bayesian Neural Network Architecture

The proposed model, **BayesianRiskNetwork**, is a Bayesian neural network designed to estimate the *unemployment duration (survival time)* and its associated *uncertainty*, given input features. The network uses **variational inference** to model probability distributions over its weights, allowing for uncertainty quantification.

Input Representation: The input consists of categorical features (Year, highest qualification, age group, and sex) encoded via embedding layers, and a continuous feature (Estimated unemployed) which is normalized. The input vector is formed by concatenating the embedded categorical features with the normalized continuous feature.

Network Structure:

- First layer: Variational linear layer with ReLU activation, batch normalization, and dropout (0.6).
- Output layer: Variational linear layer producing two outputs:
  - $\mu(x)$: predicted mean unemployment duration,
  - $\sigma(x) = \exp(\log \sigma(x))$: predicted standard deviation (uncertainty), enforced positive via exponentiation.

Concept	Mathematical Expression
Weight distribution	$q(w) = \mathcal{N}(\mu, \sigma^2)$
Sampling (Reparameterization trick)	$w = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$
Regularization (KL Divergence)	$\mathrm{KL}\big(q(w) \,\|\, p(w)\big) = \frac{1}{2} \sum_{i=1}^{n} \left(\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2\right)$

Table 2: Gaussian Weight Modeling and Regularization

Variational Inference and KL Regularization:

Loss Function (Evidence Lower Bound - ELBO): The network is trained by maximizing the Evidence Lower Bound (ELBO), combining the survival likelihood and the KL divergence regularization. The unemployment duration y is assumed to follow a **log-normal distribution**, where the network predicts the parameters of the underlying normal distribution:

$$\log y \sim \mathcal{N}(\mu, \sigma^2)$$

The log-likelihood for an observation depends on the event indicator $\delta \in \{0, 1\}$, where $\delta = 1$ for observed (uncensored) durations; $\delta = 0$ for censored durations. The log-likelihood for each data point is given by:

$$\log p(y \mid \mu, \sigma, \delta) = \delta \cdot (\log f(y)) + (1 - \delta) \cdot (\log S(y)),$$

where $f(y)$ is the probability density function (PDF) of the log-normal distribution; $S(y) = 1 - F(y)$ is the survival function, with $F(y)$ being the cumulative distribution function (CDF). Explicitly:

$$\log f(y) = \log \left(\frac{1}{y\sigma\sqrt{2\pi}} \exp \left(-\frac{(\log y - \mu)^2}{2\sigma^2} \right) \right),$$

$$\log S(y) = \log \left(1 - \Phi \left(\frac{\log y - \mu}{\sigma} \right) \right),$$

where $\Phi(\cdot)$ denotes the standard normal CDF.

The total ELBO loss combines the sum of the log-likelihood over all observations and the KL divergence term:

$$\mathcal{L}_{\text{total}} = -\sum_{i=1}^{N} (\delta_i \cdot \log f(y_i) + (1 - \delta_i) \cdot \log S(y_i)) + \beta \cdot \mathrm{KL}\big(q(w) \,\|\, p(w)\big).$$

This formulation correctly handles both observed and censored data points, making the model suitable for survival analysis with censoring. This Bayesian architecture enables the model to estimate both the expected unemployment duration and the uncertainty of these estimates, making it suitable for probabilistic survival analysis.

4.3 Training Mechanism

The Bayesian neural network model was trained using the **Adam optimizer** with a learning rate of 1×10^{-4} for a maximum of 1000 epochs. During training, the following metrics were monitored and reported: **Average training loss per epoch**, **Average KL divergence**, **KL ratio** (the ratio of KL divergence to the absolute log-likelihood), and **Validation loss**. The KL ratio helps monitor the balance between the data fit term and the regularization term.

4.4 Prediction and Survival Probability Estimation

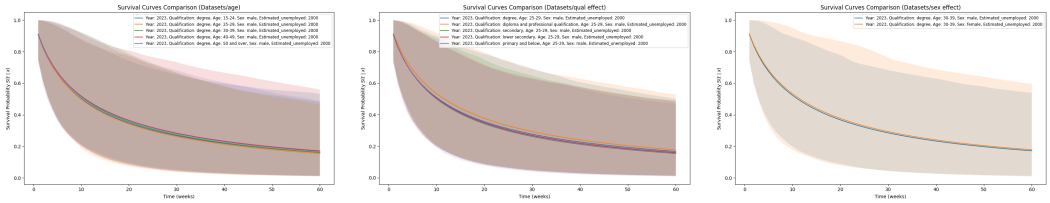
The trained Bayesian neural network estimates the *survival probability* $S(t \mid x)$, which represents the probability that an individual remains unemployed beyond time t, given their covariates x.

The network predicts two parameters for each input x:

- $\mu(x)$: mean of the log unemployment duration,
- $\sigma(x)$: standard deviation of the log unemployment duration.

Assuming the unemployment duration T follows a log-normal distribution: $\log T \sim \mathcal{N}\big(\mu(x), \sigma^2(x)\big)$. The survival function is: $S(t \mid x) = 1 - F_{\text{LogNormal}}(t; \mu(x), \sigma(x))$, where $F_{\text{LogNormal}}(t)$ is the cumulative distribution function (CDF) of the log-normal distribution. This formulation provides the probability that an individual's unemployment duration exceeds t weeks, conditioned on their features x.

4.5 Survival Analysis: Covariate Effects



(a) Age Effect (b) Qualification Effect (c) Sex Effect

Figure 6: Comparison of Survival Probability Curves by Age, Qualification, and Sex

Age Effect: People aged 20–29 with a degree have the lowest unemployment duration, while those aged 40–49 experience the highest.

Qualification Effect: Higher qualifications reduce unemployment duration, with degree holders exiting unemployment faster than lower-educated groups.

Sex Effect: Slight differences between male and female survival curves, with females showing marginally higher survival probabilities.

5 Conclusions and Future Work

References

[1] Ministry of Manpower (2024), *Unemployed Residents Aged 15 Years and Over by Age and Duration of Unemployment*, data.gov, `https://data.gov.sg/datasets/d_db95e15ceffaa368a043310479dc7d57/view`

[2] Ministry of Manpower (2024), *Unemployed Residents Aged 15 Years and Over by Highest Qualification Attained and Duration of Unemployment*, data.gov,`https://data.gov.sg/datasets/d_a0ca632fd1d6ff841f0e47298a9ab589/view`

[3] Ministry of Manpower (2024), *Median Duration of Unemployment*, data.gov, `https://data.gov.sg/datasets/d_c01a3210fb10f1a52676f97498d4ec2c/view`

[4] Ganjali, M., & Baghfalaki, T. (2012). *Bayesian Analysis of Unemployment Duration Data in the Presence of Right and Interval Censoring*. Communications in Statistics - Theory and Methods, 41(15), 2738-2751.

[5] Wong, S. C. (2018). *Transformation of Employment Patterns and Need for Career Services in Modern Singapore*. Singapore Economic Review, 63(S1), 229-246.

[6] Boškoski, P., Perne, M., Rameša, M., & Boshkoska, B. M. (2021). *Variational Bayes survival analysis for unemployment modelling* Knowledge-Based Systems, 229, 107335.

[7] Ganjali, M., Baghfalaki, T., Berridge, D., Shahid Beheshti University, Shahid Beheshti University, & Lancaster University. (2009). *A Bayesian analysis of unobserved heterogeneity for unemployment duration data in the presence of interval censoring*. In International Econometric Review (IER) (p. 25).

Appendix

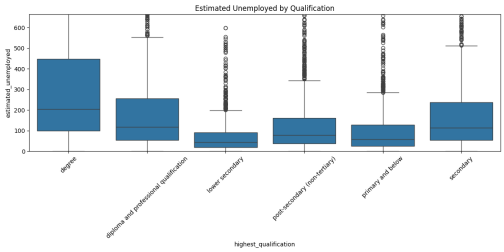


Figure 7: Estimated Number of Unemployed People by Qualification

Parameter ESS Bulk	Mean	SD	HDI 3%	HDI 97%	R-hat
alpha 2113.0	3.058	0.020	3.020	3.098	1.00
beta_year 6032.0	-0.000	0.003	-0.005	0.005	1.02
beta_sex 4650.0	0.001	0.021	-0.040	0.041	1.00
beta_age[0] 990.0	0.001	0.015	-0.032	0.031	1.01
beta_age[1] 2994.0	0.000	0.015	-0.030	0.033	1.00
beta_age[2] 2295.0	0.001	0.015	-0.032	0.030	1.00
beta_age[3] 3380.0	0.000	0.016	-0.032	0.032	1.00
beta_age[4] 1245.0	-0.000	0.015	-0.031	0.033	1.01
beta_qualification[0] 5338.0	0.001	0.013	-0.028	0.029	1.01
beta_qualification[1] 4651.0	0.000	0.014	-0.028	0.030	1.01
beta_qualification[2] 4108.0	-0.000	0.014	-0.031	0.028	1.01
beta_qualification[3] 3110.0	0.000	0.014	-0.029	0.031	1.02
beta_qualification[4] 4933.0	0.000	0.013	-0.029	0.029	1.02
beta_qualification[5] 4171.0	0.000	0.013	-0.029	0.027	1.02
sigma_age 655.0	0.016	0.018	0.000	0.046	1.01
sigma_qualification 173.0	0.014	0.013	0.001	0.038	1.03
sigma 394.0	0.764	0.009	0.747	0.781	1.01

Table 3: Bayesian Inference Summary for Model Parameters