27.04.2025

NATIONAL UNIVERSITY OF SINGAPORE

GROUP 12
SNEHA SARKAR (A0304787U)
PRERANA CHAKRABORTY (A0305018R)
LIU XIN YING (A0188304A)
LEE JIA YUN AMANDA (A0296790L)
AKSHAYA VAJPEYARR (A0307243M)

Modeling Uncertainty in Unemployment Duration

CS5340 PROJECT

TABLE OF CONTENTS

1. INTRODUCTION
2. METHODOLOGY
3. PREPROCESSING
4. BAYESIAN REGRESSION MODEL
5. SURVIVAL ANALYSIS MODEL
6. CONCLUSIONS AND FUTURE WORK
7. REFERENCES

INTRODUCTION

- **Multifaceted Nature of Unemployment Duration:** Unemployment duration is influenced by a range of personal, structural, and economic factors, making it complex and variable.

- **Importance for Stakeholders:** Understanding this complexity is crucial for policymakers, job seekers (especially students), and employers, as it helps inform decisions that affect labor market dynamics and workforce resilience.

- **Quantifying Uncertainty:** The project focuses on modeling and quantifying the uncertainty surrounding unemployment duration, providing valuable insights for proactive policy interventions and career support strategies.

- **Support for Policymakers and Job Seekers:** The model aims to help policymakers make informed decisions and assist job seekers by suggesting re-employment strategies, improving access to retraining programs, and ensuring fairer opportunities.

METHODOLOGY

- **Data Acquisition and Preprocessing**
    - Three datasets from data.gov.sg were cleaned and combined using **IPF** for reconstructing joint distribution.
- **Bayesian Regression Model**
    - Modeled uncertainty with a Student's **t-distribution likelihood.**
    - Estimated posterior distributions via **Hamiltonian Monte Carlo Variant** (NUTS).
- **Bayesian Neural Network for Survival Analysis**
    - Modeled parameter uncertainty using **variational inference** with **KL divergence** regularization.
    - Trained using **ELBO** maximization.

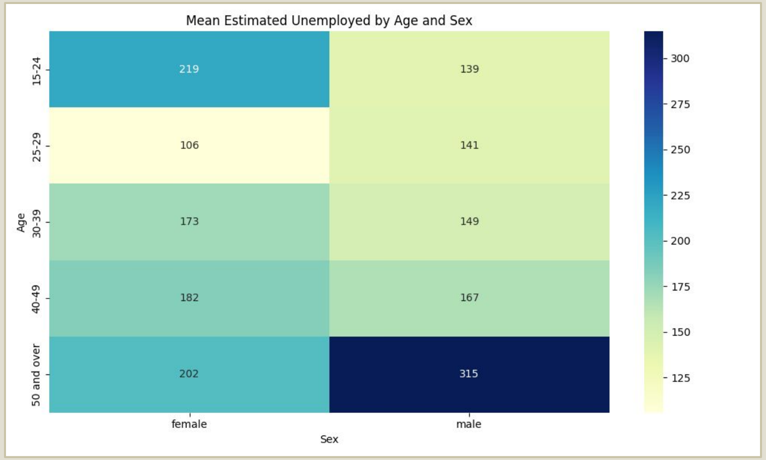
PREPROCESSING

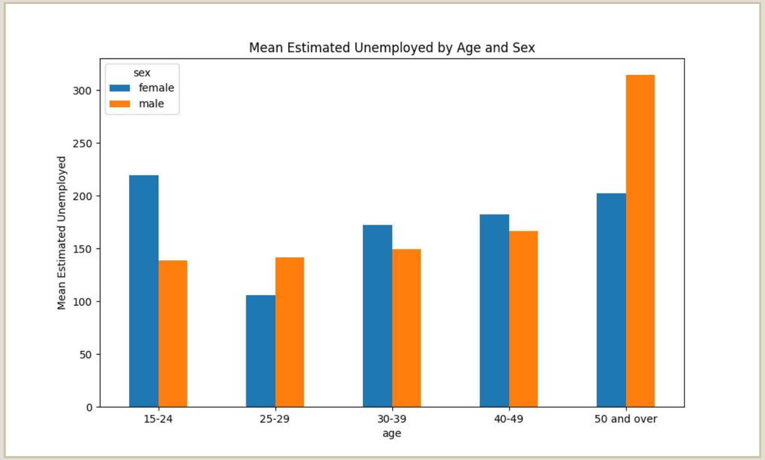
- **Official Datasets:** Three datasets from data.gov.sg, each providing valuable unemployment-related data in Singapore
    - Unemployed by Age and Duration: Distributions across different age groups and sexes.
    - Unemployed by Qualification and Duration: Distributions across qualification levels and sexes.
    - Median Duration of Unemployment: Calibration reference for median values over years.

- **Basic Cleaning:**
    - **Data Conversion:** Converted columns to numeric types, handling invalid or missing entries by assigning NaN values.
    - **Mapping Duration Ranges:** Duration ranges (e.g., "5 to 9 weeks") mapped to numeric midpoints for easier modeling.
    - **Handling Missing Data:** Careful imputation to avoid bias and ensure valid input for further steps.

PREPROCESSING

- **Marginal Computations:**
  - Computation of Marginals: For each year, computed marginal counts for:
    - Age × Sex from the age-duration dataset.
    - Qualification × Sex from the qualification-duration dataset.
  - Purpose: Marginal probability distributions used for joint probability table reconstruction.





PREPROCESSING

Iterative Proportional Fitting (IPF):

- IPF Algorithm: Used to estimate the joint distribution of age, sex, and qualification by adjusting the initial uniform guess to match observed marginals.
- Convergence: Algorithm iteratively refines the joint distribution until it aligns with the marginal distributions.

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

Duration Probability Estimation:

- Age-conditioned probability:

$$P(duration = d \mid age = a,\, sex = s,\, year = y)$$

- Qualification-conditioned probability:

$$P(duration = d \mid qualification = q,\, sex = s,\, year = y)$$

- Imputation: Missing or sparse values were carefully imputed to stabilize the model.

PREPROCESSING

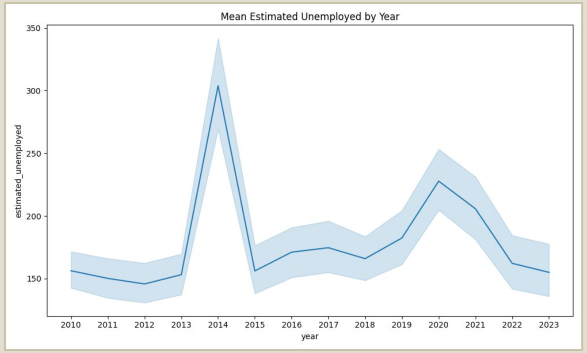
Hybrid Probability Calculation: The hybrid probability was obtained by linearly combining the age-conditioned and qualification-conditioned probabilities with equal weighting (w= 0.5), reflecting no strong prior on which factor is more important

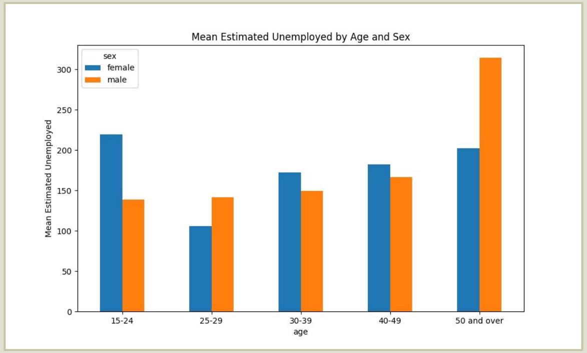
$$P_{hybrid}(d) = 0.5 \times P_{age}(d) + 0.5 \times P_{qual}(d)$$

Final Estimation and Validation:

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

The resulting estimated unemployed counts were rounded to the nearest integer to produce a final dataset suitable for modeling.





BAYESIAN REGRESSION MODEL

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

Response Variable: The midpoint of each duration interval (e.g., "10 to 14" → 12) is used as the observed duration. The response is then log-transformed:

$$y_i = \log(duration_i)$$

Feature encoding: The input features were processed as follows:

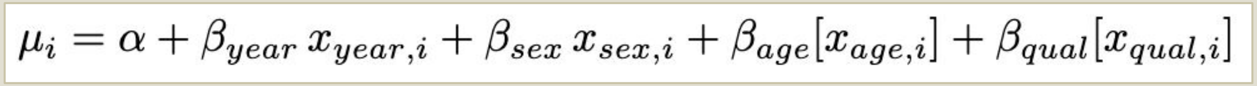
- *Year (centered):* Each year's value was centered by subtracting the mean year across the dataset:

$$x_{year,i} = year_i - \overline{year}$$

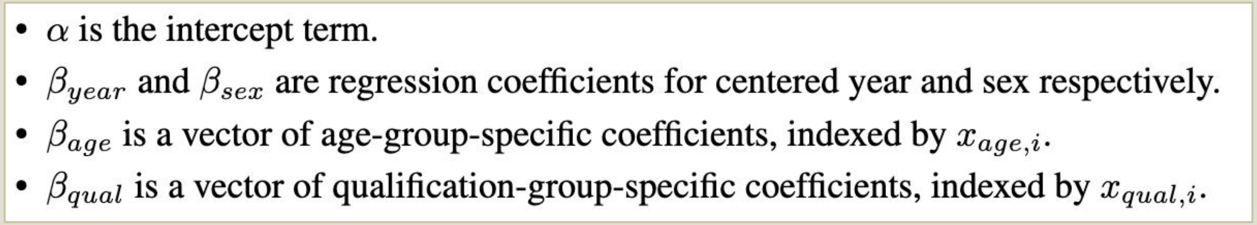
- *Sex (binary encoded):* Encoded as 0 for male and 1 for female.
- *Age group (categorical encoded):* Each unique age group (e.g., 15–24, 25–29) was mapped to an integer index.
- *Highest qualification (categorical encoded):* Each qualification category (e.g., degree,diploma, secondary) was mapped to an integer index.

BAYESIAN REGRESSION MODEL

Hierarchical Linear Predictor - The expected log-duration for individual i is modeled as:

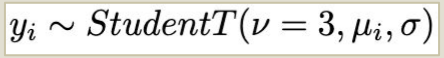


$$\mu_i = \alpha + \beta_{year}\, x_{year,i} + \beta_{sex}\, x_{sex,i} + \beta_{age}[x_{age,i}] + \beta_{qual}[x_{qual,i}]$$



- 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 $x_{age,i}$.
- β_{qual} is a vector of qualification-group-specific coefficients, indexed by $x_{qual,i}$.

Likelihood: The observed log-duration y_i for each individual i is modeled using a Student’s t-distribution, which is robust to outliers:



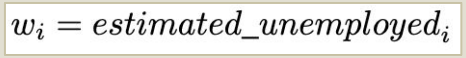
$$y_i \sim StudentT(\nu = 3, \mu_i, \sigma)$$

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

BAYESIAN REGRESSION MODEL

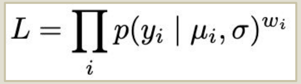
Likelihood (continued):

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



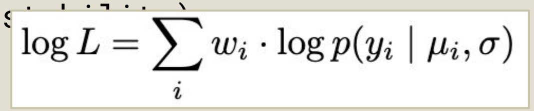
$$w_i = estimated\_unemployed_i$$

- The weighted likelihood across all individuals is given by:



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

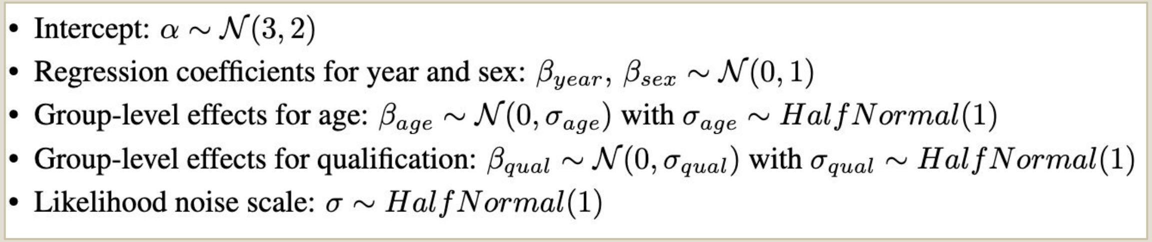
- Taking the log of the likelihood (to improve numerical stability), we obtain the weighted log-likelihood:



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

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:

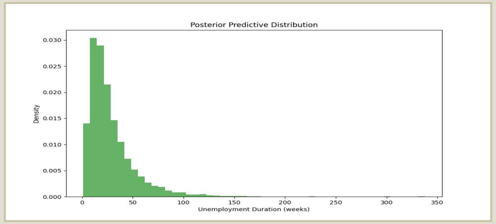


- Intercept: $\alpha \sim \mathcal{N}(3, 2)$
- Regression coefficients for year and sex: $\beta_{year}, \beta_{sex} \sim \mathcal{N}(0, 1)$
- Group-level effects for age: $\beta_{age} \sim \mathcal{N}(0, \sigma_{age})$ with $\sigma_{age} \sim HalfNormal(1)$
- Group-level effects for qualification: $\beta_{qual}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

NATIONAL UNIVERSITY OF SINGAPORE

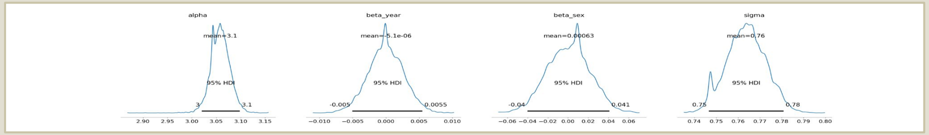
27.04.2025

BAYESIAN REGRESSION MODEL



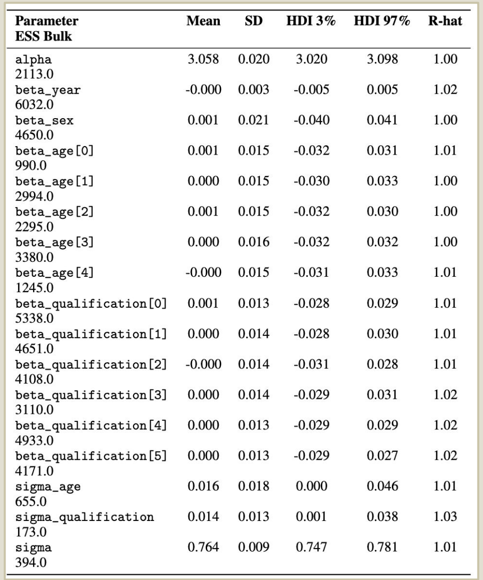
Model Fitting and Posterior Inference

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.



Prediction for New Cases

- **Inputs**: Year, sex, age group, and qualification (e.g., "degree" for a female in the 30-39 age group, year 2025).
- **Process**:
  - Normalize Year: Subtract the mean year from the training data.
  - Encode Categorical Variables: Find integer indices for age group and qualification.
  - Compute Linear Predictor: Use posterior samples to calculate a linear predictor.
  - Generate Prediction: Sample from log-normal distribution to obtain prediction.
- **Results**:
  - Median Prediction: 21.59 weeks.
  - 95% Credible Interval: 4.87 to 93.66 weeks (reflecting uncertainty).



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

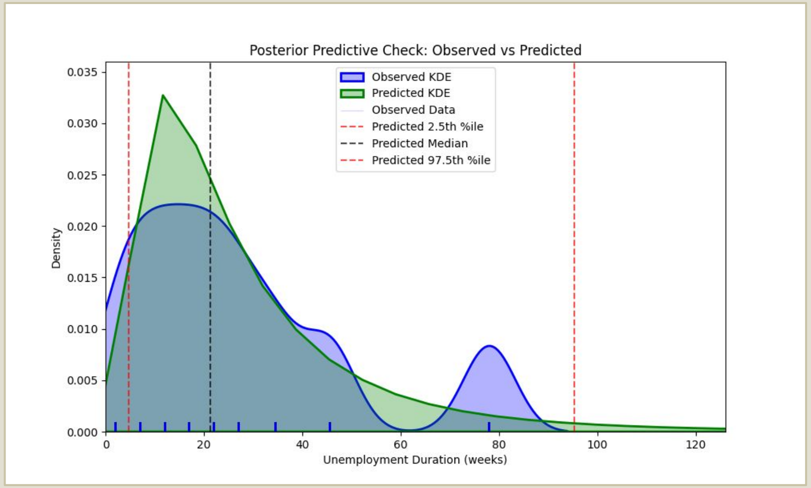
BAYESIAN REGRESSION MODEL

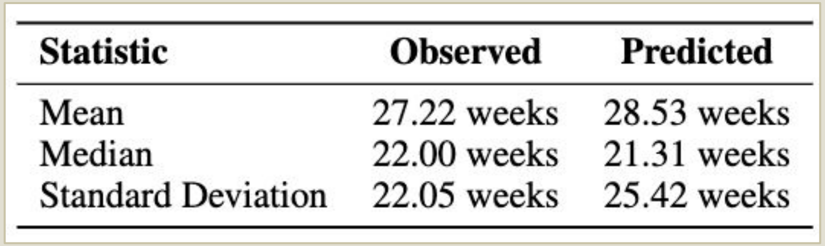
Model Diagnostics & Visualization

- **WAIC** (8000 posterior samples, 7560 observations):
  - elpd_waic = -11056.77 (SE = 76.20)
  - p_waic = 6.39
- **LOO** (8000 posterior samples, 7560 observations):
  - elpd_loo = -11056.77 (SE = 76.20)
  - p_loo = 6.39
- **Pareto k** Diagnostic:
  - Good: 99.9% of values (7552) ≤ 0.70
  - Bad/Very Bad: 0.1% of values (>0.70)
- **Results**: Low WAIC and LOO scores with no problematic Pareto k values suggest good model fit and inference quality.
- **Posterior Predictive Checks**: Predicted vs. observed distributions of unemployment duration were compared using summary stats and kernel density estimates (KDEs).

Technical Notes

- Censoring - Interval-censored data is used, the likelihood has been adapted to account for lower and upper bounds.
- Numerical Stability - Log-transformations and clipping are used to avoid issues with zero or negative durations.



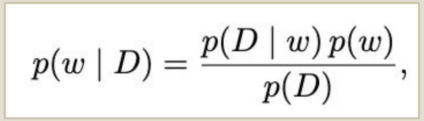


Statistic	Observed	Predicted
Mean	27.22 weeks	28.53 weeks
Median	22.00 weeks	21.31 weeks
Standard Deviation	22.05 weeks	25.42 weeks

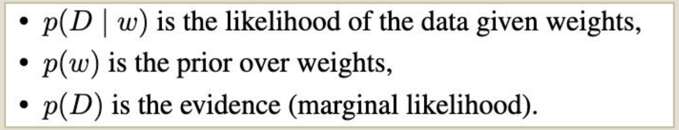
SURVIVAL ANALYSIS MODEL

Bayesian Modeling Approach

The goal is to infer the posterior distribution over weights p(w | D)



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



- $p(D \mid w)$ 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 | x).

SURVIVAL ANALYSIS MODEL

Bayesian 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

- **Categorical features**: Year, highest qualification, age group, and sex — encoded via embedding layers.
- **Continuous feature**: Estimated unemployed — normalized.

The input vector is formed by concatenating the embedded categorical features with the normalized continuous features.

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.

SURVIVAL ANALYSIS MODEL

Bayesian Network Architecture

Variational Inference and KL Regularization: Each weight w is modeled as a Gaussian distribution .Each linear layer is parameterized by a mean and log-variance, which are learned during training. The variational layers use a reparameterization trick for sampling:

$$q(w) = \mathcal{N}(\mu, \sigma^2).$$

$$W = \mu_W + \sigma_W \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

$$\mathrm{KL}\big(q(w) \,\|\, p(w)\big) = \frac{1}{2} \sum_{i=1}^{n} \left(\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2\right).$$

Regularization is achieved by minimizing the Kullback-Leibler (KL) divergence between the approximate posterior and a standard normal prior.

The unemployment duration y is assumed to follow a log-normal distribution, where the network predicts the parameters of the underlying normal distribution:

The log-likelihood for an observation depends on the event indicator δ ∊ {0, 1}, where: $\log y \sim \mathcal{N}(\mu, \sigma^2).$

- δ = 1 for observed (uncensored) durations,
- δ = 0 for censored durations.

The log-likelihood for each data point is given by:

- 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).

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

SURVIVAL ANALYSIS MODEL

Bayesian Network Architecture

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 Φ(·) denotes the standard normal CDF.

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

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

This formulation correctly handles both observed and censored data points, making the model suitable for survival analysis with censoring. 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

SURVIVAL ANALYSIS MODEL

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.

Prediction and Survival Probability Estimation

The trained Bayesian neural network estimates the survival probability S(t | 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:

- μ(x): mean of the log unemployment duration,
- σ(x): standard deviation of the log unemployment duration.

Assuming the unemployment duration T follows a log-normal distribution:

$$\log T \sim \mathcal{N}\left(\mu(x), \sigma^2(x)\right),$$

the survival function is:

$$S(t \mid x) = 1 - F_{LogNormal}(t; \mu(x), \sigma(x)),$$

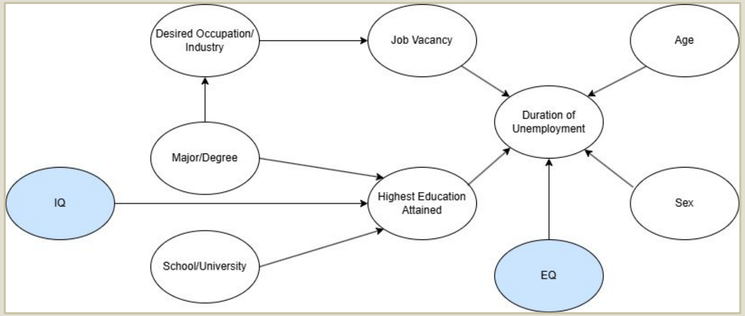
where $F_{LogNormal}$(t) is the cumulative distribution function (CDF) of the log-normal distribution. This formulation provides the probability that an individual’s unemployment duration exceeds t weeks, conditioned on their features x.

CONCLUSION

- Our Bayesian framework successfully models uncertainty in unemployment duration across demographic and educational factors. The hierarchical model showed strong performance metrics (elpd_waic = -11056.77) with 99.9% of Pareto k values in the “good” range. Posterior predictive checks confirmed the model’s accuracy, with predicted distributions closely matching observed data (observed median: 22.00 weeks; predicted median: 21.31 weeks).
- The model’s ability to provide credible intervals (e.g., 4.87 to 93.66 weeks for our example case) offers valuable uncertainty quantification for both individual career planning and policy formulation.
- Survival analysis revealed that age significantly impacts unemployment duration, with individuals aged 20-29 experiencing the shortest unemployment periods while those aged 40-49 face longer durations. Higher education levels demonstrated a clear inverse relationship with unemployment duration. Sex differences were detected but relatively modest.

FUTURE WORK

- **Economic Context Integration:** Incorporating macroeconomic indicators to disentangle personal versus structural unemployment factors.
- **Causal Analysis:** Exploring intervention effectiveness across different population segments using causal inference techniques.



- **Applied Tools:** Developing decision support systems that provide personalized unemployment forecasts and targeted re-employment strategies.
- **Comprehensive Bayesian Network:** Developing an advanced model incorporating both observable factors (job vacancies, education) and latent variables (IQ, EQ) to better predict unemployment duration and enable more targeted interventions.

These extensions would enhance both the predictive power of our model and its practical utility for policymakers and job seekers navigating uncertain labor markets.

THANK YOU