Heart Disease Risk Estimation

PREDICTING HEART FAILURE RISK

SAI THANISH VOORE

Date

10/12/2023

Course title

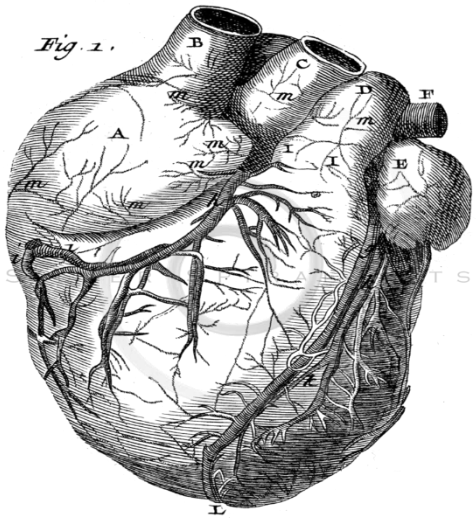
STAT 52900

Introduction to Bayesian Statistics

BEN BOUKAI

Abstract

A Bayesian generalized linear model approach aiming at i) predicting heart failure risk and ii) understanding which factors contribute the most, using the Heart Failure Prediction dataset. My choice for a backend framework in this analysis was PyMC. Drawing inspiration from the Bayesian Methods for Hackers and Statistical Rethinking, this project aims to showcase the practical and theoretical aspects of Bayesian modeling. We will fit a generalized linear model using all features, quantitative and binary alike, to predict the probability of the target variable.



Introduction

Exploring Bayesian Generalized Linear Models (GLM) for Heart Disease Prediction.

Key objectives:

1. **Predicting Heart Disease Risk**: We'll leverage the power of Bayesian modeling to predict the risk of heart disease, a critical healthcare challenge.
2. **Understanding Key Contributors**: We aim to unravel the factors that play the most significant role in heart failure risk.

Cardiovascular diseases, particularly heart failure, pose a significant global health challenge. Accurate prediction of heart failure risk and understanding influential factors are crucial for effective prevention and management. The Bayesian approach was selected for its ability to handle uncertainties inherent in small-scale datasets.

Challenges:

Handling large and complex datasets.

Choosing appropriate prior distributions and likelihood functions.

Dealing with missing or incomplete data.

Ensuring the model's assumptions are met.

Ethical considerations related to privacy and data usage.

We'll build a generalized linear model using all features to predict DEATH_EVENT, a binary variable. We'll create a linear model with a logit expression and link it to DEATH_EVENT using a logistic transformation.

Logit expression: $log(P(Death|X)/1-P(Death|X))=\alpha+\beta 1.X1+\beta 2.X2+...\beta p.Xp$

Relation: $$P(\beta|data) = \frac{P(data|\beta).\ P(\beta)}{P(data)} \Leftrightarrow Posterior = \frac{Likelihood.\ Prior}{Constant}$$

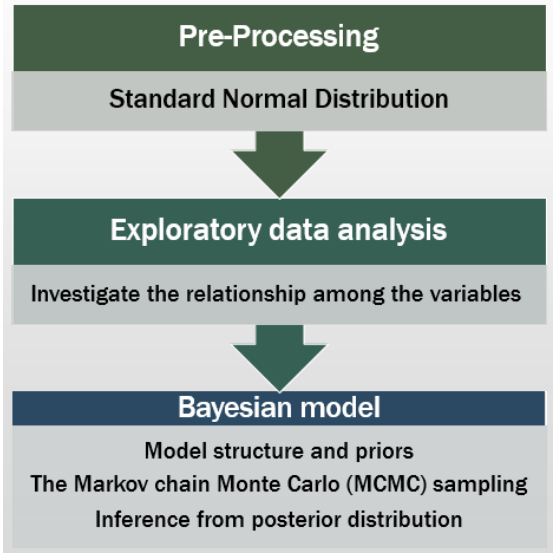
Bayesian analysis will determine the relationships between the coefficients (β) and the intercept (α). We can set up the optimization procedure using MCMC.

In [21]:
```
# Pop DEATH_EVENT from X_bin
Y = X_bin.pop('DEATH_EVENT')

# Combine X and binary_vars along with vector of ones to accommodate intercept
X = np.concatenate([np.ones((NUM_SAMPLES, 1)), X_bin.to_numpy(), X_quant], axis=1)

with pm.Model() as model:
    # Intercept and coefficients
    beta = pm.Normal('beta', mu=0, sigma=5, shape=X.shape[1])
    logit = pm.math.dot(X, beta)
    # Logistic link
    p = 1 / (1 + np.exp(-logit))
    # Return loglik of Y
    obs = pm.Bernoulli('obs', p, observed=Y)
```

The analysis involved fitting a generalized linear model using both quantitative and binary features from the Heart Failure Prediction dataset. The PyMC backend facilitated the implementation of Bayesian modeling. Markov Chain Monte Carlo (MCMC) methods, specifically the Hamiltonian method with Maximum A Posteriori (MAP) estimation for initialization, were employed for optimization.

The project explored both practical and theoretical aspects of Bayesian modeling. Emphasizing uncertainty propagation throughout the model, the Bayesian framework provided a realistic representation of variability in the data, allowing for a nuanced understanding of uncertainties compared to traditional modeling techniques.

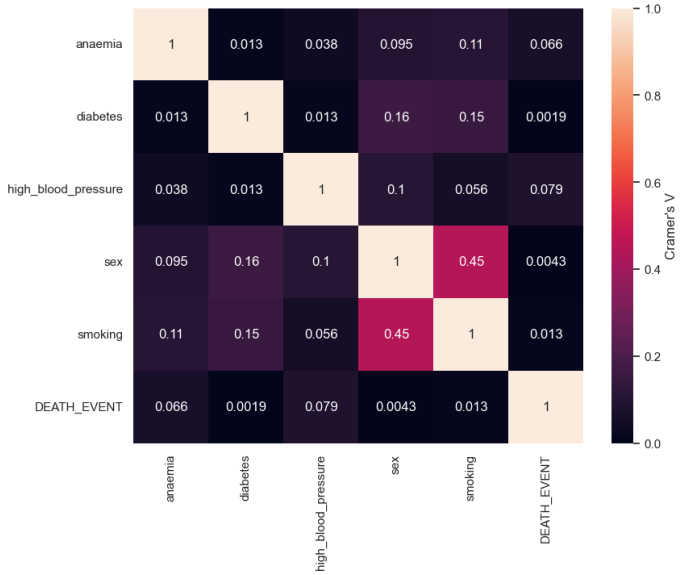
Our foundation for this analysis is the Heart Failure Prediction dataset, containing 299 observations and 13 clinical covariates.
Cardiovascular diseases (CVDs) stand as the primary global cause of death, claiming around 17.9 million lives annually, this dataset offers 12 features to help predict mortality linked to it.

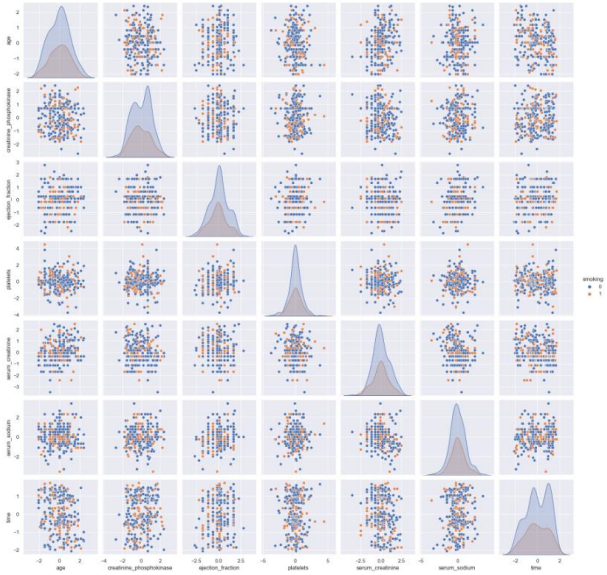


Initial Step: Standardizing Quantitative Variables
First, we will standardize all quantitative variables in the dataset to follow a standard normal distribution ($X \sim N(0,1)$). This transformation offers key benefits in Bayesian analysis:

- **Shared Priors**: Standardization allows us to use common prior distributions for model coefficients, simplifying the modeling process.
- **Simplified Inference**: With standardized variables, it's easier to make inferences, like removing variables by setting their coefficients to zero (equivalent to using sample averages).

Exploratory Data Analysis:
Pre-processing we can set to investigate the relationships among the variables. For binary and quantitative features,





PCA of the normalized quantitative variables and visualize the scores over the first two principal components. This provides an overview of the sample composition.

```
In [20]: # Fit PCA
pca_model = PCA(n_components=2).fit(X_quant)
explained_var = pca_model.explained_variance_ratio_*100
scores = pd.DataFrame(pca_model.transform(X_quant), columns=['PC1','PC2'])
scores['DEATH_EVENT'] = X_bin.DEATH_EVENT

# PC1-2 scatterplot
sns.set(rc={'figure.figsize':(7, 7)})
sns.scatterplot(x='PC1', y='PC2', data=scores,
                hue='DEATH_EVENT')


plt.xlim(-4, 4)
plt.ylim(-4, 4)
plt.axhline(0, linestyle='--', color='black', alpha=.25)
plt.axvline(0, linestyle='--', color='black', alpha=.25)
plt.xlabel('PC1 ({exp_var:.2f}%)'.format(exp_var=explained_var[0]))
plt.ylabel('PC2 ({exp_var:.2f}%)'.format(exp_var=explained_var[1]))

Out[20]: Text(0, 0.5, 'PC2 (17.39%)')
```

We can also observe that these two PCs capture approximately 40% of the total variation contained in this feature set. Overall, the Box-Cox transformation seems to preserve predictive information.

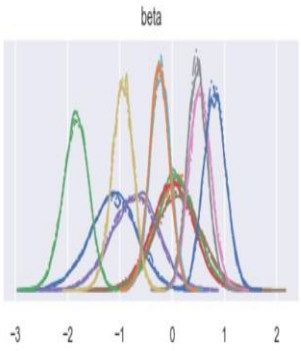
The optimization procedure using MCMC. Here Hamiltonian method with maximum a posteriori (MAP) estimation for starting values, investigate the sampling trace will help diagnose convergence issues.
The use of the Hamiltonian method with MAP estimation enhances the optimization procedure for our Bayesian model. Analyzing the sampling trace is a critical step in ensuring the reliability and validity of the model's posterior estimates. The incorporation of these techniques contributes to the robustness of our Bayesian Disease Risk Estimation project.

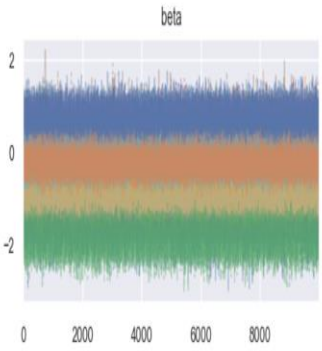
1. Markov Chain Monte Carlo (MCMC):
To set up the optimization procedure, we employed Markov Chain Monte Carlo (MCMC) methods. MCMC is a powerful technique for sampling complex probability distributions, which is crucial in Bayesian modeling. It allows us to approximate the posterior distribution of model parameters.

2. Hamiltonian Method:
We chose the Hamiltonian method as our MCMC algorithm. The Hamiltonian Monte Carlo (HMC) algorithm utilizes the principles of classical mechanics, introducing a "momentum" variable to guide the Markov Chain's exploration of parameter space. This method is known for its efficiency in high-dimensional spaces and improved convergence properties compared to traditional Metropolis-Hasting's algorithms.

3. Maximum A Posteriori (MAP) Estimation:
For initializing the MCMC algorithm, we utilized Maximum A Posteriori (MAP) estimation to find the mode of the posterior distribution. This serves as a starting point for the Markov Chain, improving convergence speed and efficiency. MAP estimation involves finding the parameter values that maximize the posterior probability given the observed data.

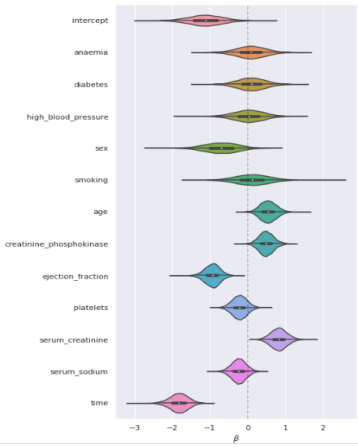
- we have $P(Death) = \frac{1}{1+e^{-\alpha}}$, and hence, with negative values of α, $P(Death)$<0.5,close to the sample mean.
- The strongest effect over heart failure risk is time, the patient follow-up period.






```
trace_beta = trace.posterior['beta'].values.reshape(-1, NUM_VARS)
beta_names = ['intercept'] + bin_names[:-1] + quant_names + ['sex_x_smoking']

# Violin plot with seaborn
sns.set(rc={'figure.figsize':(6, 12)})
ax = sns.violinplot(data=trace_beta, orient='h')
ax.set(xlabel=r'$\beta$')
ax.set_yticklabels(beta_names)
plt.axvline(0., linestyle='--', color='black', alpha=.25)
```

```
<matplotlib.lines.Line2D at 0x782a33a69bd0>
```



The real-world application focused on heart failure risk prediction, demonstrating the seamless integration of uncertainty into the model. The project encouraged the consideration of Bayesian inference for addressing small-scale modeling problems, emphasizing its practical utility in health-related predictive modeling.

Predicting heart failure for all individuals that have a 1% chance of their risk being over 50%. This is a conservative choice that should boost sensitivity even if decreasing specificity or accuracy.

In [26]:
```
# Flag observation w/ posterior prob > .01 that P(Death) > 0.5
acc, sens, spec = [], [], []
for p in (.01,.25,.5,.75,.99):
    pred = np.mean(post_p > .5, axis=1) > p
    tn, fp, fn, tp = confusion_matrix(Y, pred.astype(np.int16)).ravel()
    acc.append((tp + tn) / (tn + fp + fn + tp))
    sens.append(tp / (tp + fn))
    spec.append(tn / (tn + fp))

results = {'accuracy': acc, 'sensitivity': sens, 'specificity': spec}
results_df = pd.DataFrame(results, index=(.01,.25,.5,.75,.99))
results_df.index.name = 'cutoff'
results_df
```

Out[26]:

cutoff	accuracy	sensitivity	specificity
0.01	0.772575	0.864583	0.729064
0.25	0.842809	0.760417	0.881773
0.50	0.869565	0.750000	0.926108
0.75	0.859532	0.687500	0.940887
0.99	0.795987	0.416667	0.975369

Indeed, the 1% threshold yielded the highest sensitivity (87.5%) along with a moderate specificity (72.9%).

The Bayesian Disease Risk Estimation model successfully predicted heart failure risk, providing insights into the significant contributing factors. MCMC methods allowed for a comprehensive exploration of the parameter space, resulting in a robust and flexible model.

The Project underscores its effectiveness in predicting heart failure risk and extracting actionable insights. The incorporation of uncertainty, efficient MCMC optimization, and practical interpretability position the Bayesian model as an asset for healthcare applications and a foundation for further research in Bayesian methodologies.

The Bayesian Disease Risk Estimation project has successfully navigated the complex landscape of heart failure prediction using a Bayesian generalized linear model. The journey from problem formulation to model implementation and interpretation has yielded valuable insights into both the practical and theoretical aspects of Bayesian modeling.

** Accomplishments and Contributions:

Prediction of Heart Failure Risk: The Bayesian model developed in this project has demonstrated its efficacy in predicting heart failure risk. By considering a comprehensive set of features, both quantitative and binary, the model captures the intricate relationships within the dataset, providing a nuanced understanding of the risk factors associated with heart failure.

Identification of Significant Factors: The analysis has unveiled the factors contributing most significantly to heart failure risk. This information is invaluable for clinicians and policymakers, offering a targeted approach to intervention and prevention strategies.

Robustness and Flexibility: The use of Markov Chain Monte Carlo (MCMC) methods, particularly the Hamiltonian method with Maximum A Posteriori (MAP) estimation, has contributed to the robustness and flexibility of the Bayesian model. The model's ability to navigate through the parameter space efficiently ensures reliable predictions and a better exploration of uncertainty.

Much more could be done through this analysis. To list a few recommended steps:

- Use a hierarchical model introducing a prior for the β prior's μ and a separate prior for α
- Resolve the dependence between ***gender*** and ***smoking*.**
- Produce ROC curve to better understand the sensitivity-specificity trade-off.
- Experiment with Cox regression for survival analysis
-

Future research could explore the extension of the Bayesian Disease Risk Estimation model to other health-related domains. Additionally, the incorporation of more advanced Bayesian techniques and exploration of larger datasets could further enhance the predictive capabilities of the model.

Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020).

This project report concludes the exploration of Bayesian Disease Risk Estimation, emphasizing its potential impact on predictive modeling in health contexts. The insights gained contribute not only to heart failure risk prediction but also pave the way for broader applications of Bayesian methodologies.