Bayesian Statistics Homework 1
Sebastijan Trojer

Introduction

In this homework we aimed to model life expectancy using multiple linear regression. The main questions were whether life expectancy (LE) is improving over time, how are GDP per capita and population correlated with LE, comparing LE across continents and estimating our own LE and comparing it to that of the professor.

Methods

Data

The data consists of the target variable $y \in \mathbb{R}$ and 4 predictors - year of birth $\in \mathbb{R}$, continent $\in$ {Africa, Americas, Asia, Europe, Oceania}, GDP per capita $\in \mathbb{R}$ and population of residing country $\in \mathbb{R}$. The dataset consists of 1704 samples.

Modeling

We modeled the data using multiple linear regression, which we defined as $y \sim N(X\beta, \sigma^2)$ where the first column of X is a column of ones, modeling the intercept. Since the numerical features have vastly different scales, year, GDP per capita and population variables were normalized to be in [0, 1] range. Since these variables seemingly have no direct relationships, we were able to rescale them independently without losing information. Next step was to encode the continent variable, since it is categorical and therefore we have to use dummy variables. Africa was picked as reference category. Lastly we prepared the data matrix and fit the model using STAN and R.

In cases where we wanted to compare individual draws instead of means, the linear combinations XB were computed, and samples for comparison were drawn from $N(X\beta, \sigma^2)$. Standard error and other statistics were computed using the *mcse* function. Default priors were used. Parameters were sampled from 4 parallel chains.

Results

The MCMC diagnostics implies no reason for concern, the traceplots indicate proper sampling, $\hat{r}$ is 1 for all variables and *ESS* ranges from 2800 to 3400.

Life expectancy trend

To determine the trend in life expectancy, we investigated β_{year}. Initially we computed the mean,, standard error and 95% confidence interval (CI) of the raw β_{year} and the model is 100% certain that LE is rising over time. To get more tangible results, we rescaled the β_{year} to the original scale, and the results show that each year, LE rises by 0.286 ± 0.000 years, that is approximately 100 days. Figure 2 shows the values of the fitted coefficients, that shows that all of them are positive.

Effect of GDP per capita and population

Next we investigated the correlation of GDP and population to LE. First we computed the correlation between the feature and target variables, getting the correlation coefficient 0.584 and 0.065 for GDP and population, respectively. We also analyzed the posterior distributions of the regression coefficients. The posterior mean of β_{GDP} was positive (33.90 ± 0.04), with CI entirely above zero, indicating a positive association between GDP per capita and LE. The posterior mean of $\beta_{population}$ was also positive (2.42 ± 0.03); however, its CI partially overlapped negative values, suggesting less certain evidence for a positive association. Furthermore, we found that the correlation of these two features is actually slightly negative.

Inter-continent life expectancy analysis

In this section we investigate how likely an average european born in 2001 is to live longer than people from other continents, born in 2001, and how likely that scenario is for individuals. First we aggregate the data and compute mean feature values per continent within 1999 and 2003. Then we compute probabilities for average individuals per continent, and compare means of the distributions. Table 1 shows the probabilities and difference in expected LE between EU and other continents. To compare individual people's LE, we sample the posterior and compute means and standard errors. The results are in Table 1 and the posteriors are shown in Figure 1.

	Africa	Americas	Asia	Oceania
Mean continent values				
Probability (EU > Other) Difference (years)	1 ± 0 25.1	1 ± 0 8.8	1 ± 0 13.3	0.028 ± 0.003 -2.7
Posterior samples				
Probability (EU > Other) Difference (years)	0.99 ± 0.001 24.8	0.81 ± 0.006 8.7	0.91 ± 0.005 13.1	0.38 ± 0.008 -3.0

Table 1. Comparison of life expectancy between Europe and other continents. Probabilities that Europeans live longer are on top; expected differences (years) are below. Top block: mean continent values; bottom block: posterior samples.

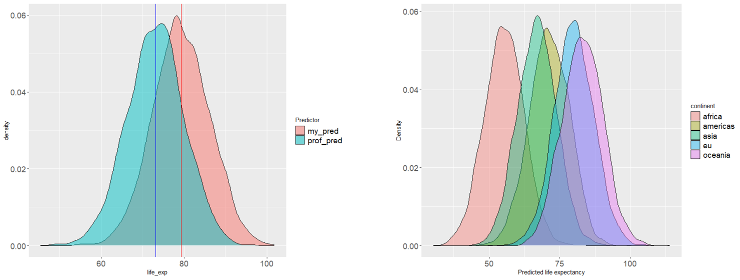

Figure 1. Posterior distribution of Slovenian from 2001 and 1985 (left) and posterior distributions of people from 2001 from different continents (right).

Predicting and comparing our life expectancies

The next topic was inferring our own life expectancy. To ensure as faithful predictions as possible, we used Slovenia GDP and population data from 2001, which were 18.765€ (PPP, as described for the values in the dataset) and approximately 2 million. The new feature vector was normalized in the same fashion as the original data, and used to get the vector of predicted means. Then, based on these means, samples were drawn from the normal distribution. The mean and 95% CIs are reported in Table 2. If instead of real data, we use mean feature data from all European entries, the predictions change up to a decimal point. Next we also computed the samples for the professor. By searching for information about GDP and population on the web, we found approximate entries of 13.481€ GDP per capita and population of about 50 thousand less than in 2001. After sampling from both distributions, the probability of me living longer is 74.4%. More details are shown in Table 2. On average, someone born in Slovenia in 2001 is expected to live 6.3 years longer than someone born in 1985. Posteriors of 1985 and 2001 Slovenian are shown in Figure 1.

Subject	Mean LE	95% CI
Me (2001)	79.5	[66.5, 92.8]
Professor (1985)	73.2	[60.0, 86.8]
P(Me > Professor) = 0.73 ± 0.007		

Table 2. Predicted life expectancy (years). CIs of the difference between samples indicate that I might die somewhere between 14 years before up to 24 years after the professor.

Discussion

We successfully modeled life expectation and answered several more nuanced questions about it. However our approach has some limitations, specifically, it assumes linear relationships, which is not very realistic (higher GDP doesn't necessarily mean longer life). It also assumes that life expectancy changes at a constant rate, which ignores certain events in real life. Another assumption is that the entries in the dataset are independent, with no colinearity, which we'd have to address with a suitable prior (regularization). Another problem is that the data is outdated, and the model would benefit from more novel data, especially given how fast modern world is changing. In some experiments we also made some decisions that may have influenced the results, such as what feature values to use, when comparing people across continents etc.

Appendix

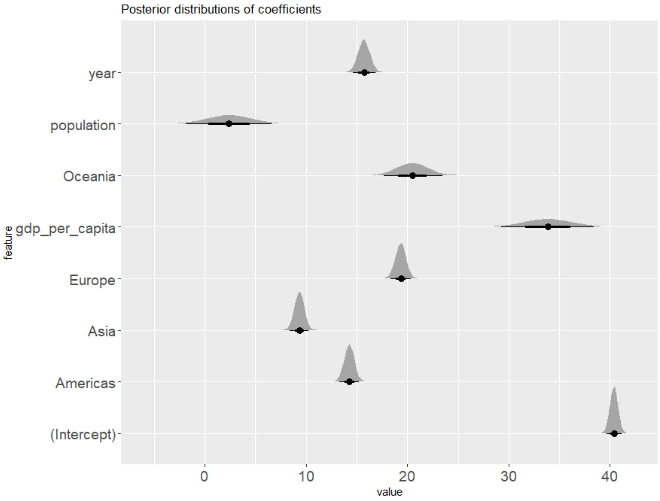

Figure 2. Coefficient distributions. Notice that all the coefficients, with the exception of $\beta_{population}$ are strictly positive, implying a positive impact on LE. Keep in mind that the continent coefficients are relative to Africa, which has the shortest LE.