Data Analysis Project

Coursera
2023

1 Introduction

This project aims to predict the progression of pulmonary fibrosis, a complex lung disease with no known cause or cure. We employ hierarchical Bayesian statistics and data analysis techniques to examine lung scans, basic patient data, and initial lung function tests for making these predictions. The goal is to provide valuable insights into the disease's course and potentially accelerate the development of new treatments for pulmonary fibrosis.

2 Data Analysis

The FVC dataset consists of medical information related to patients with pulmonary fibrosis. It includes data on the forced vital capacity (FVC) of the patients, which measures the volume of air they can exhale. This dataset contains records of patients' FVC measurements taken at various time points over the course of approximately 1-2 years. It provides baseline chest CT scans, clinical information, and FVC measurements. In the training set, the complete history of FVC measurements is available, while in the test set, only the initial FVC measurement is provided. The goal is to predict the final three FVC measurements for each patient and provide a confidence value for those predictions. Additionally, the dataset includes information such as patient identifiers, the timing of measurements relative to the baseline CT scan, age, sex, and smoking status. Fig. 1 shows the dataframe structure and a few lines of data.

	Patient	Weeks	FVC	Percent	Age	Sex	SmokingStatus
	<chr>	<int>	<int>	<dbl>	<int>	<chr>	<chr>
1	ID00007637202177411956430	-4	2315	58.25365	79	Male	Ex-smoker
2	ID00007637202177411956430	5	2214	55.71213	79	Male	Ex-smoker
3	ID00007637202177411956430	7	2061	51.86210	79	Male	Ex-smoker
4	ID00007637202177411956430	9	2144	53.95068	79	Male	Ex-smoker
5	ID00007637202177411956430	11	2069	52.06341	79	Male	Ex-smoker
6	ID00007637202177411956430	17	2101	52.86865	79	Male	Ex-smoker
7	ID00007637202177411956430	29	2000	50.32713	79	Male	Ex-smoker
8	ID00007637202177411956430	41	2064	51.93759	79	Male	Ex-smoker
9	ID00007637202177411956430	57	2057	51.76145	79	Male	Ex-smoker
10	ID00009637202177434476278	8	3660	85.28288	69	Male	Ex-smoker

Figure 1: FVC dataset.

3 Model

In this model, we utilize Bayesian statistics to predict lung function (FVC) changes over time for individual patients. We employ a linear regression framework with patient-specific intercepts and slopes. Each patient's FVC trajectory is modeled as a linear function, allowing for variability in their starting point (intercept) and rate of change (slope).

We treat patient-specific intercepts and slopes as random variables with normal distributions, capturing the uncertainty and variation among patients. We use hyperparameters to control the spread of these patient-specific parameters, offering flexibility in modeling variations. Precision terms, inverses of variances, help fine-tune the distributions of intercepts and slopes. The standard deviation represents the spread of actual FVC measurements. The model is described as follows

$$FVC_i \sim \mathcal{N}(\alpha_i + \beta_i * \mathrm{W}eeks_i, \sigma)$$

$$\tau = \Gamma(0.001, 0.001)$$

$$\alpha_i \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha)$$
$$\beta_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$$

$$\mu_\alpha \sim \mathcal{N}(0, 500)$$
$$\tau_\alpha \sim \Gamma(0.001, 0.001)$$

$$\mu_\beta \sim \mathcal{N}(0, 3)$$
$$\tau_\beta \sim \Gamma(0.001, 0.001)$$

$$\sigma = 1/\sqrt{\tau}$$
$$\sigma_\alpha = 1/\sqrt{\tau_\alpha}$$
$$\sigma_\beta = 1/\sqrt{\tau_\beta}.$$

In Fig. 1, we observe the standard linear regression applied to three selected patients. However, traditional linear regression falls short when we have limited data for each patient, and the disease effects tend to be quite similar across individuals. This is where hierarchical Bayesian statistics shine. By recognizing the commonalities in disease effects among patients, hierarchical Bayesian modeling effectively addresses the data scarcity issue, offering a more robust and insightful approach.

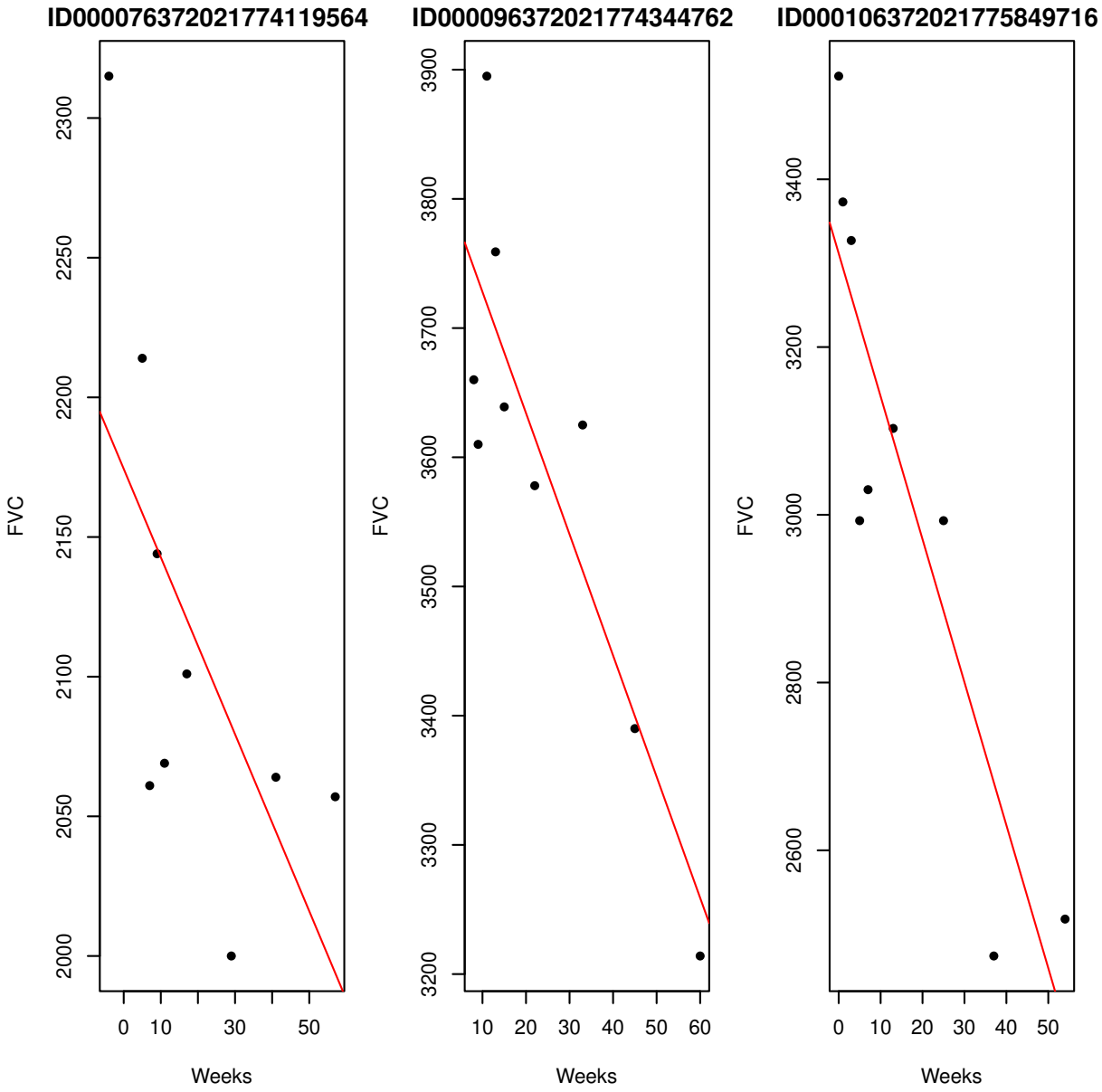

Figure 2: Standard linear regression for three patients.

Below is the code for the Jags model

```
// RJAGS Model
" model {
    for (i in 1:N){
        FVC[i] ~ dnorm(alpha[patient_code[i]] + beta[patient_code[i]]*Weeks[i],
            tau)

    }

    for (i in 1:n_patients){
        alpha[i] ~ dnorm(mu_a, tau_a)
        beta[i] ~ dnorm(mu_b, tau_b)
    }
```

```
    sigma <- 1/sqrt(tau)
    tau ~ dgamma(1.0E-3, 1.0E-3)

    mu_a ~ dnorm(0, 1/(500*500))
    tau_a ~ dgamma(1.0E-3, 1.0E-3)
    sigma_a <- 1/sqrt(tau_a)

    mu_b ~ dnorm(0, 1/(3*3))
    tau_b ~ dgamma(1.0E-3, 1.0E-3)
    sigma_b <- 1/sqrt(tau_b)

} "
```

4 Results

The model ran for 11000 iterations where the first 1000 were discarded. Fig. 3 shows the sampled posterior distributions for α, β, and σ while Fig. 4 shows for μ_α, μ_β, σ_α, and σ_β, respectively.

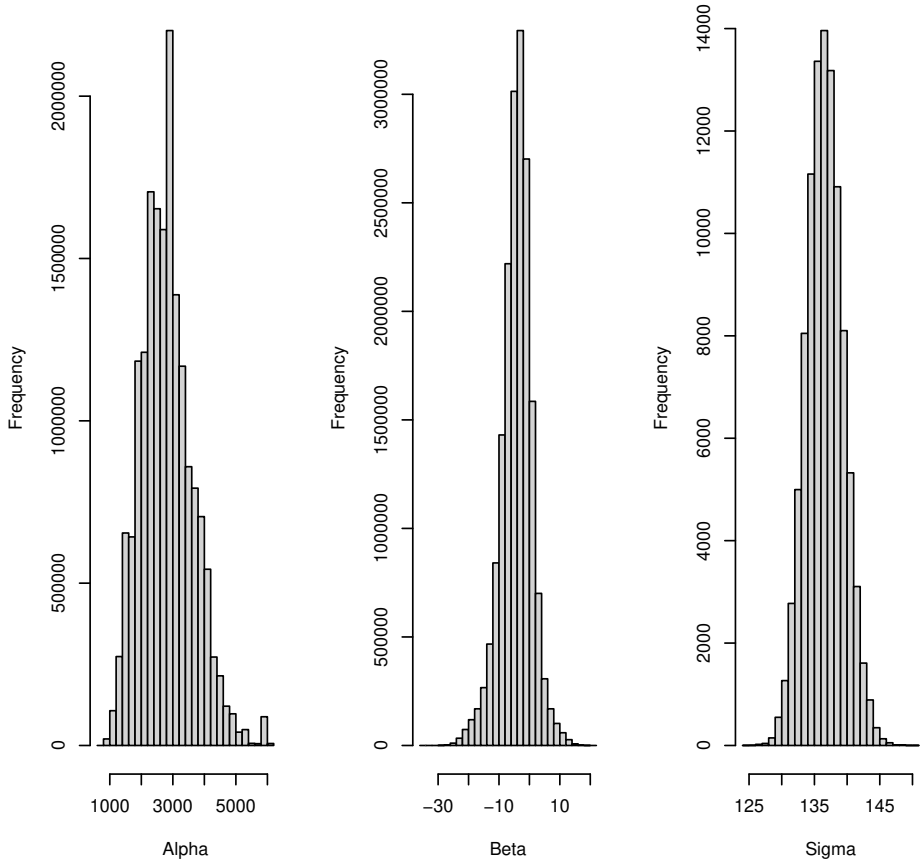

Figure 3: Sampled posterior distributions for α, β, and σ.

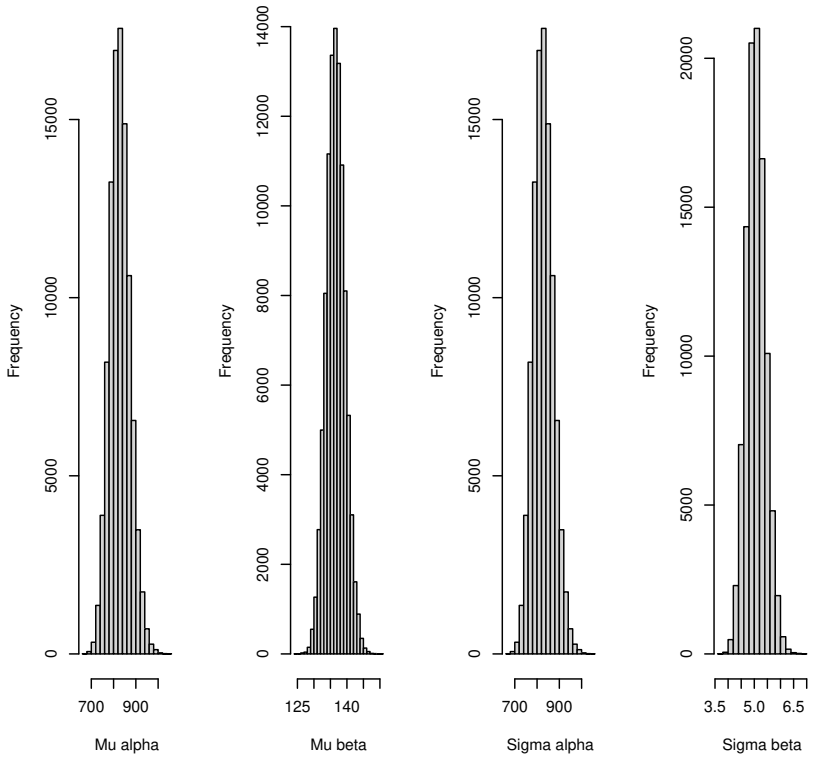

Figure 4: Sampled posterior distributions for μ_α, μ_β, σ_α, andσ_β.

The average intercept is $\alpha \approx 2809.23$, while the average slope is $\beta \approx -4.25$. Now that we have the intercept and slope for every patient, we can try to make some predictions as shown in Fig. 5.

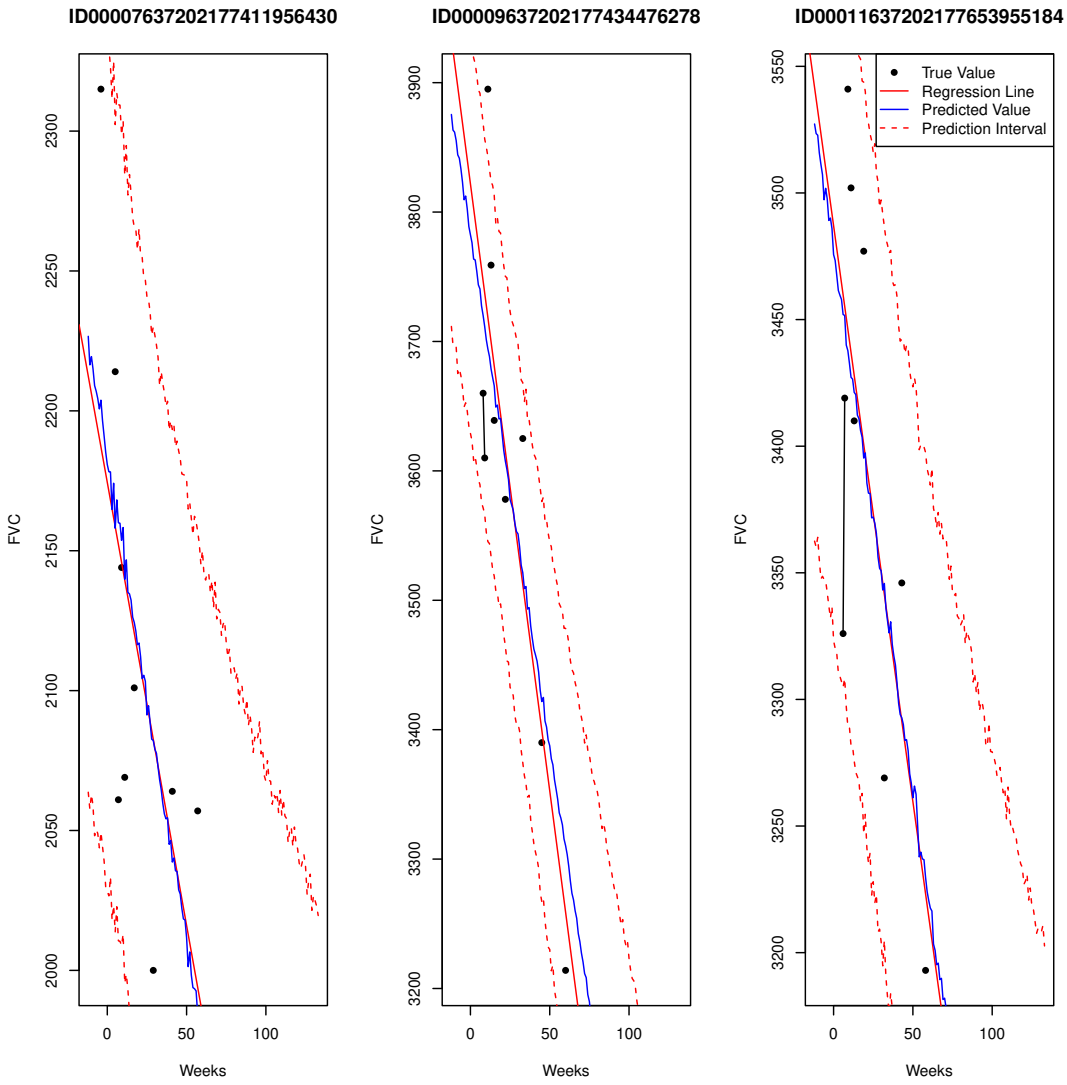

Figure 5: Bayesian model predictions.

5 Conclusions

Our Bayesian Linear Regression model has learned to predict FVC effectively. It closely aligns with the deterministic linear regression, as seen in the close resemblance between the red line and the blue line. More importantly, the model excels in predicting uncertainty, as indicated by the dashed red lines, which stand one standard deviation above and below the mean FVC line.

The model tends to predict higher uncertainty in cases where data points are scattered, such as the 1st and 3rd patients. Conversely, when data points are closely grouped, like the 2nd patient, the model confidently predicts a narrower region between the red dashed lines.

Furthermore, across all patients, a noticeable trend emerges: as we project further into the future (increasing the number of weeks), the uncertainty widens. This reflects the model's recognition of increasing unpredictability over time.