Bayesian Logistic Regression Modelling for Back Pain Abnormality

Numaer Zaker <nzaker3@gatech.edu>
IsYE 6420 - Bayesian Statistics - Fall 2020 Project
Georgia Institute of Technology

1 Introduction

Back pain is a health condition that affects millions of people globally. Unfortunately, determining the causes of back pain is challenging due to the structural complexity of the spine. However, modern technology allows us to quantify characteristics of an individual's spine. With Bayesian logistic regression modelling, it's possible to process spine data to diagnose back pain abnormalities. The benefits of producing this model are:

1. Proactively inform a patient of potential future back pain and advise on preventative steps (e.g. keep lumbar at 180 degrees at all time)

2. Diagnose patients with chronic existing back pain and prescribe treatment (pelvic joint inflammation medication)

3. Understand key drivers behind a normal healthy back and an abnormal unhealthy back.

In this paper, we use Bayesian logistic regression methodology to capture the relationships between back pain abnormality and back features. There are several reasons for this choice over other statistical learning models:

1. Given that our patient spine dataset is not large (310 data points), Bayesian methodology works elegantly by using expert opinion (priors) and updating it through the dataset (likelihood). The end result is a posterior distribution that incorporates the best of expert opinion and observed patient spine data.

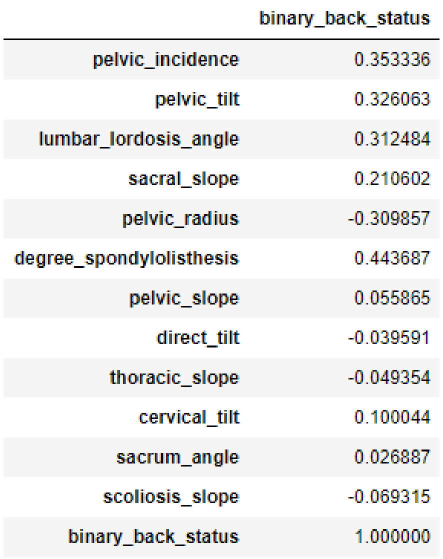
2. Using Bayesian methods allow us to use probability distributions and credible sets for modelling features (β coefficients). This gives us the advantage of understanding how confident we are with our estimates and how they are associated with back pain abnormalities (response variable).

In this paper, we first analyze the spine data and detail all the characteristics (# features, # rows, collinearity, etc). Then, we describe our methodology for creating a Bayesian logistic regression model between spine features and back pain abnormality. Lastly, we conclude our paper with results and key findings.

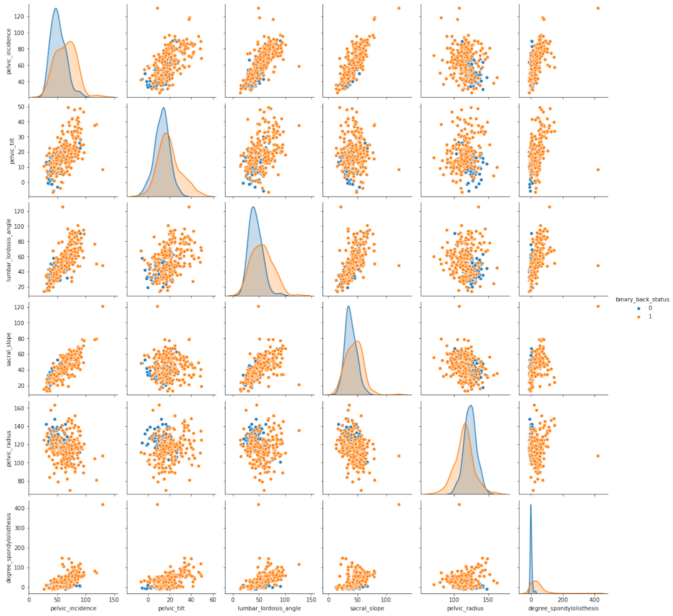
2 Spine Data Exploration

For this project, we use the "Lower Back Pain Symptoms Dataset" [1] uploaded on Kaggle.com. Our patient spine dataset has 310 data points, where each data point contains 12 features that describe various metrics around an individual's back. These twelve features are: **pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, degree spondylolithesis, pelvic slope, direct tilt, thoracic slope, cervival tilt, sacrum angle, scolosis slope**. Our 13th column is the status of the back which is a binary value of "Abnormal" or "Normal". We'll assume that each of these data points are independent and identically distributed (i.i.d).

We remove the features not strongly correlated with our response variable (assumptions of regression). We also use a correlation scatterplot to help remove features with high collinearity (little additive explanatory power, more model complexity). Below are the graphs (left is correlation between feature and response, right is scatter correlation):



	binary_back_status
pelvic_incidence	0.353336
pelvic_tilt	0.326063
lumbar_lordosis_angle	0.312484
sacral_slope	0.210602
pelvic_radius	-0.309857
degree_spondylolisthesis	0.443687
pelvic_slope	0.055865
direct_tilt	-0.039591
thoracic_slope	-0.049354
cervical_tilt	0.100044
sacrum_angle	0.026887
scoliosis_slope	-0.069315
binary_back_status	1.000000



In the left figure, we observe that pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle, scoliosis slope are features that have very weak correlation with binary back status (abnormality). We prune these features out for the rest of the analysis. In the figure on the right, we color-encode each of the datapoints to be either abnormal (blue) or normal (orange) to see if any these remaining columns shows patterns of separating the clusters. We look at **pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral**

slope, pelvic radius, degree spondylolithesis. These distributions resemble a normal distribution with varying degrees of skew and kurtosis. Normal backs tend to have positive kurtosis, while abnormal backs tend to have less kurtosis but more skew. The scatter plots show clear separation of points between abnormal and normal backs, making them very viable candidates for our model features. We select the features that do not display patterns of multicollinearity and in this case it is we see that pelvic incidence, lumbar lordosis angle, and sacral slope show positive linear correlation. For the rest of our analysis, we focus on the following columns: **pelvic tilt, lumbar lordosis angle, pelvic radius, and degree spondylolisthesis**. These columns exhibit the ideal characteristics for logistic regression (little collinearity, independent observations, correlated with response).

3 Bayesian Modelling and Spine Abnormality Classification

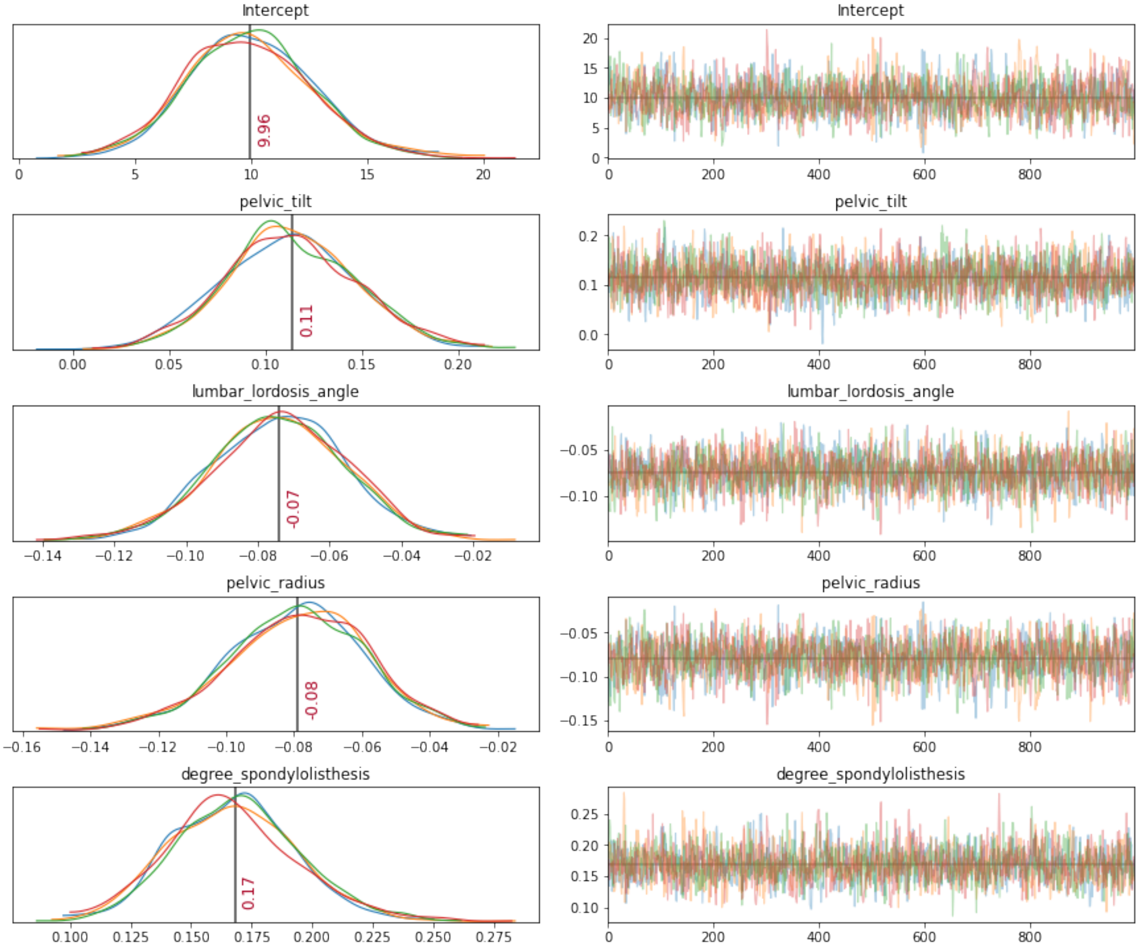
The Bayesian logistic regression model performs a regression between spine abnormality (1 = abnormal, 0 = normal) and the selected features **pelvic tilt, lumbar lordosis angle, pelvic radius, and degree spondylolisthesis**. We formulate our Bayesian logistic regression model based off of professor's "Engineering Biostatistics" logit section [2] below:

$$
\begin{gathered}
y_i \sim Bin(n_i, p_i) \\
\text{logit}(p_i) = \beta_0 + \beta_1(\text{pelvic\_tilt}) + \beta_2(\text{lumbar\_lordosis\_angle}) + \\
\beta_3(\text{pelvic\_radius}) + \beta_4(\text{degree\_spondylolisthesis})
\end{gathered} \tag{1}
$$

Where y is the number of abnormal spines, p is the probability of having an abnormal spine, and n, is the the total number of spine data points. We look to produce posterior distributions for each of our β parameters from our β priors multiplied by the likelihood function $L(y_i|\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, x)$. We use non-informative priors for our β random variables: $\pi(\beta) = N(0, 10^{-6})$[3] with precision parameter (an expert may use informative ones). Our sampler samples from the posterior distribution with the kernel shown below:

$$
\begin{gathered}
\pi(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4|x, y) \propto L(y|\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, x)\pi(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) \\
\pi(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4|x, y) \propto \prod_{i=1}^{n} (\tfrac{1}{1+e^{-y_i}})^{y}(1 - \tfrac{1}{1+e^{-y_i}})^{1-y_i}\tfrac{1}{\sigma\sqrt{2\pi}}e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}
\end{gathered} \tag{2}
$$

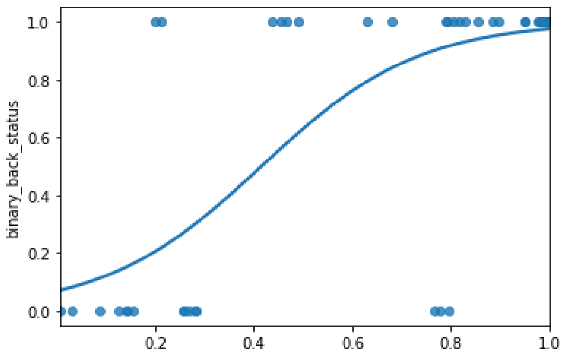
In class, we used Metropolis-Hastings and Gibbs samplers for sampling. In this project, we use PyMC3's default No U-Turn sampler (NUTS) [4] which converges faster than MH and Gibbs samplers. For training our Bayesian regression model, we split the dataset into training (80%) and testing (20%). Our Bayesian model trains on 248 data points. After sampling 1000 times, we arrive with the following posterior distributions for our β values:

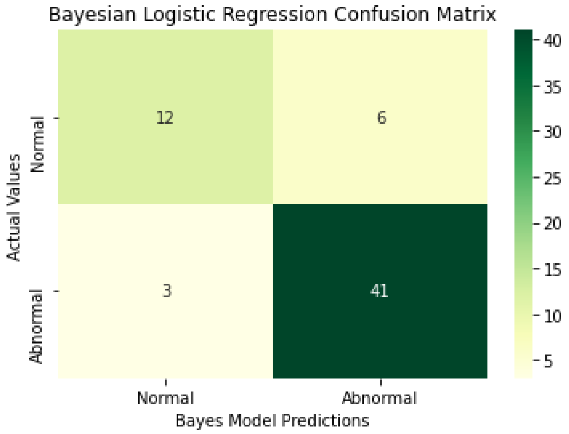


On the left, we see that the posterior is in the same form as our prior distributions but with the means and variance updated from the likelihood functions. On the right, we observe our sample trace path which shows convergence and proper behavior. From the above posteriors, we can deduce the equation that calculates the probability a person has an abnormal back:

$$
\begin{aligned}
p &= \frac{1}{1+e^{-(\beta_0+\beta_1(\text{pelvic\_tilt})+\beta_2(\text{lumbar\_lordosis\_angle})+\beta_3(\text{pelvic\_radius})+\beta_4(\text{degree\_spondylolisthesis}))}} \\
&= \frac{1}{1+e^{-(9.96+0.11(\text{pelvic\_tilt})+0.07(\text{lumbar\_lordosis\_angle})+-0.08(\text{pelvic\_radius})+0.17(\text{degree\_spondylolisthesis}))}}
\end{aligned} \tag{3}
$$

To test the performance of our model, we pass the remaining 62 test spine data points into our Bayesian logistic regression model. It classifies a back as abnormal if $p > 0.5$ and normal if $p \leq 0.5$ (an expert may want to adjust this). The figure on the left below shows the true classification (y-axis) against the model probability (x-axis). The figure on the right is the confusion matrix, where we see 53 points were correctly classified and 9 were wrong. Our Bayesian logistic regression model boasts a strong prediction accuracy of 85.48%.





4 Findings & Conclusion

From our model, we observe that patients with an abnormal back tend to have a smaller lumbar lordosis angle and smaller pelvic radius. But they also tend to have higher degrees of pelvic tilt and spondylolisthesis. The spondylolithesis degree seems to be the strongest indicator of an unhealthy back (coefficient of 0.17). Bayesian logistic regression modelling is shown here to be more intuitive than the traditional frequentist approach. For each of our β predictors, we are given credible set intervals and means that allow us to diagnose patients with certainty levels as opposed to frequentist absolutes. Our Bayesian model performance is also a testament to show the method works very well for small data sets even when our priors are non-informative. We've only touched the surface in this paper, but Bayesian statistics has far more applications in the medical space.

References

[1] Lower Back Pain Symptoms Dataset. Sammy123, 2016
`https://www.kaggle.com/sammy123/lower-back-pain-symptoms-dataset`

[2] Engineering Biostatistics: an Introduction Using MATLAB and WinBUGS
Brani Vidakovic, Wiley, 2017

[3] PyMC3 ”Generalized Linear Models”
`https://docs.pymc.io/api/glm.html`

[4] PyMC3 ”NUTS Sampling”
`https://docs.pymc.io/api/inference.html`

[5] PyMC3 API
`https://docs.pymc.io/api.html`