**Body Fat Prediction Using Bayesian Regression**

Sankhadip Roy

Department of Computer Science and Statistics

University of Rhode Island, Kingston

**Introduction:**

In this report we will fit a Bayesian Linear Regression Model to predict the percentage of body fat. Body fat, one measure of health, has been accurately estimated by an underwater weighing technique. But fitting body fat to the other measurements using multiple regression provides a convenient way of estimating body fat for men using only a scale and a measuring tape. We will be using a dataset in which the percentage of body fat, age, weight, height, and ten body circumference measurements (e.g., abdomen) were recorded for 252 men. The data were generously supplied by Dr. A. Garth Fisher, Human Performance Research Center, Brigham Young University, Provo, Utah 84602, who gave permission to freely distribute the data and use them for non-commercial purposes. Reference to the data is made in Penrose, et al. [1].

The remainder of the article is organized as follows. At first, we describe the data mentioning all the variables with possible information associated with them. Then we summarize the data checking some immediate statistics, scatter plots, correlation matrix etc. After that we introduce a suitable linear regression model with bodyfat as response against all other variables except density. Lastly, we apply Bayesian inference towards those parameters and dispersion of the linear regression model.

**Data Description:**

The dataset has 252 observations on the following 15 variables.

Table

Description automatically generated

**Table 1**

Bodyfat: Percentage of bodyfat using Brozek's equation [2], 457/Density - 414.2, Age: Age of the men(yrs),

Weight in lbs, Height in inches, Neck, Chest, Abdomen, Hip, Ankle, Thigh, Biceps(expanded), Forearm, Wrist (all circumferences) in cm and Density in gm/cubic cm. We can clearly see density in Brozek’s formula for percentage of bodyfat. So, we remove density variable from the dataset and keep the rest.

**Summary of the dataset:**

We can check the quick summary statistics (Table 2), correlation matrix (Table 3) and scatter plot of all variables against bodyfat (Fig 1) in this page.

|  |  |
| --- | --- |
| Table  Description automatically generated  **Table 2** | A picture containing table  Description automatically generated  **Table 3** |

Diagram

Description automatically generated

**Fig 1**

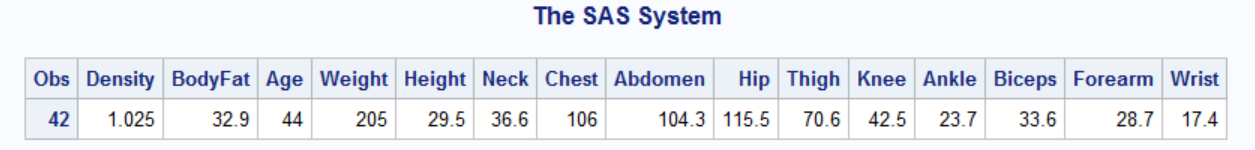
From the summary statistics we do not see existence of many outliers. From the correlation matrix, we can see that bodyfat has strong positive correlations with abdomen whereas weight has quite strong correlation with neck, chest, abdomen, hip, thigh knee and biceps. Scatter plots are mostly following linear pattern. But we can identify a huge outlier with the plot between height and bodyfat. We can further confirm that using the histogram of height.

Chart, histogram

Description automatically generated

**Fig 2**

The histogram of height (Fig 2) shows existence of data(height) less than 40 inches which does not look normal. So, we check all the data (Table 3) where height is less than 40 inches.

 Chart, histogram

Description automatically generated

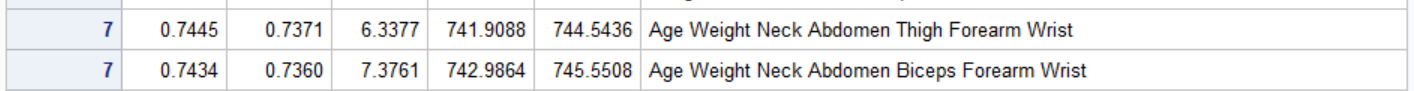
**Table 3 Fig 3**

We can see an observation where height is 29.5 inches, but the weight is 205 Ibs which does not seem possible. So, we replace the 29.5 by our natural guess 69.5. Fig 3 is the histogram after making that change.

**Building Linear Regression Model:**

We have tried to fit best possible linear regression model for bodyfat against all other variables (except density) by using exhaustive search method and ended up with the following choices for the best possible values of Adjusted R-Square, C(p), AIC and BIC respectively. We have decided to take the second choice.





**Table 4**

After regressing bodyfat against age, weight, neck, abdomen, biceps, forearm, and wrist we get the following estimates with VIF value.

|  |  |
| --- | --- |
| Table  Description automatically generated | Table  Description automatically generated  Table  Description automatically generated |

**Table 5**

From Table 5 we can see there are few p-values higher than 0.05. But we have kept those variables as those variables are important. Besides, vif value for weight is 12.57 which is because weight has strong correlation with all the other variables in the model. But we need to keep weight in the model.

**Diagnostic Plots:**

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Fig 4**

All the residual plot looks (Fig 4) quite random. The Q-Q plot has very small tails which can be ignored. Besides, the numbers of influential points are low. Over all the diagnostic plots are quite favorable for fitting linear regression model. One last thing we check the existence of **heteroscedasticity**.

**BP test:**

Text

Description automatically generated with medium confidence

**Table 6**

This test has a p-value much more than that of a significance level of 0.05. Therefore, we infer that **heteroscedasticity** is not present. Now we can apply Bayesian inferences on the parameters of our regression model.

**Bayesian Regression:**

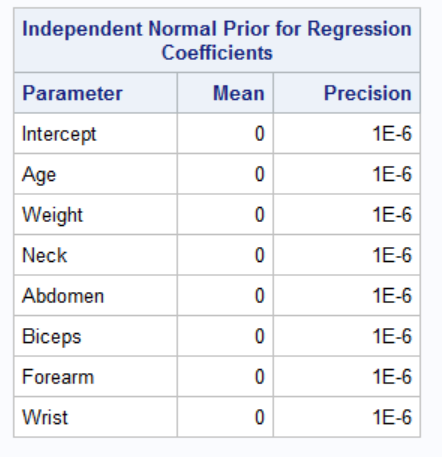
We will use weakly informative prior with mean=0 and var=10^6 for all the regression parameters.

Table

Description automatically generatedTable

Description automatically generated

**Table 7**

Graphical user interface, application

Description automatically generated with medium confidence

**Table 8**

|  |  |  |
| --- | --- | --- |
|  | Graphical user interface, chart  Description automatically generated | Chart, histogram  Description automatically generated |
| Chart  Description automatically generated | Chart, histogram  Description automatically generated | Chart, histogram  Description automatically generated with medium confidence |
| Chart, histogram  Description automatically generated | Chart  Description automatically generated | Chart  Description automatically generated |

**Fig 5**

Table 7 and 8 give us the posterior summary, quantile and HPD interval of parameters. The values are close to what we see in the original regression model. Fig 5 shows us all the trace plot, autocorrelation, and posterior densities of all the parameters including the dispersion. The fig 5 shows that the MCMC

simulations converged with proper mixing. There is burning of 1000 simulations as instructed in the code (Appendix). The DIC is 1460.27.

**Training and Testing the model:**

We have divided the data set in 70:30. Then we have trained the data with train data set and finally test it with test data set. We get the following result with R-square value of the regular linear model.

|  |  |
| --- | --- |
| RSME | R-square |
| 4.2301 | 0.7434 |

**Conclusion:**

We have applied linear regression to predict the percentage of bodyfat using the bodyfat dataset. Our model satisfies the assumptions of Linear Regression Model except few irregularities which were unavoidable. Later, we have used Bayesian inference for the parameters of our model along with the dispersion with weakly informative prior. The MCMC simulations converged successfully. The RMSE error is 4.2301 with 74.34% accountability of the model. Further investigations can be done to find a more effective model to increase the R-Square value and to decrease the RMSE. The little VIF problem can possibly be resolved by centralizing the data by mean. We can also try more informative priors. But the convergence of the MCMC process must be secured.

**References:**

1. Penrose, K., Nelson, A., and Fisher, A. (1985), "Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques" (abstract), \_Medicine and Science in Sports and Exercise\_,17(2), 189.
2. Brozek, Grande, Anderson, and Keys (1963, p. 137)