Analysis of the causes of diabetes by Bayesian regression

Minxuan Chen, Ruixuan Deng and Yishan Zhang
University of Michigan

December 15, 2023

Abstract

This project presents an innovative approach to understanding the causes of diabetes through Bayesian regression analysis. Our team at the University of Michigan has developed a robust model that adeptly handles the multi-level variability inherent in complex medical datasets. We commence our analysis with the OLS method, refined through model selection guided by the AIC. Combining what we have learned in the course, we use bayesian regression which choose loosely constrained prior distributions and contrust hierarchical model, enhancing the model's simplicity and applicability. Our work predominantly features the use of Stan for Bayesian linear regression, evolving the model into a hierarchical framework. This project is a unique blend of traditional regression techniques and advanced Bayesian methods. It provides key insights into the effective selection and estimation of model parameters, significantly contributing to our understanding of diabetes causation. Through this research, we offer a comprehensive, statistically sound framework for analyzing complex health-related data, ultimately aiding in better disease understanding and management.

1 Introduction

Many models of great interest in Mathematics and Statistics have been motivated by the goal of having a goodness of fit and accurate predictions for future values. Approaches like Forward Selection, Backward Elimination, and All Subsets regression, along with their various combinations, are commonly employed to craft "good" linear models that predict a response variable y based on measured covariates $x_1, x_2, \cdot, x_m$[2, 4, 3]. The essence of "goodness" in these models is often defined in terms of prediction accuracy. Our objective is to diligently enhance this goodness.

Upon delving into Bayesian linear regression, we've discovered that the Bayesian framework allows for the intuitive integration of prior knowledge into the model.

This becomes particularly advantageous when dealing with limited data, providing a pathway to more robust estimates. Unlike traditional regression methods, which often rely on point estimates, Bayesian methods offer a nuanced approach. They enable a delicate balance in estimates by incorporating prior knowledge, proving especially beneficial in scenarios with limited data availability.

In addition,bayesian regression provides a natural way to perform hyperparameter tuning. The choice of prior distributions can be seen as beliefs about hyperparameters, reducing the reliance on domain-specific knowledge in many cases. So in our project, we introduce the basic bayesian regression analysis and bayesian regression analysis with hyperparameter, then compare them with tradition OLS regression analysis finding which could fit the data best.

1.1 Our contributions

In this project, we introduce a comprehensive Bayesian hierarchical model, a particularly valuable tool when faced with data exhibiting variability across different levels, demanding a flexible modeling of uncertainties within these levels. The distinctive feature of Bayesian hierarchical models lies in their multi-layered structure, allowing for the representation of intricate relationships within the data.

Initially, we employ traditional Ordinary Least Squares (OLS) regression to analyze the data. The results are then subjected to the Akaike Information Criterion (AIC) for model selection, incorporating a combination of backward elimination and Forward Selection. Overall, we find this model to be robust, effectively explaining the underlying patterns in the data.

Furthermore, we carefully select a prior distribution for the coefficients, encompassing the slope of each variable and the intercept in the model. This prior distribution encapsulates our beliefs about these parameters before observing the data. Given the model's simplicity and limited information about the parameters, we opt for a loose prior without specific preferences. Subsequently, we leverage Stan to facilitate Bayesian linear regression, yielding insightful conclusions.

Lastly, by imposing a uniform distribution for the priors, we transform the entire Bayesian linear regression model into a hierarchical model. This hierarchical model is then applied to conduct Bayesian regression, leading to comprehensive and informative conclusions.

1.2 Results

In our project, we use OLS regression, bayesian regression and hierarchical model to analyze the diabetes data, but what was unexpected was that we obtained similar conclusions under the three different analysis methods.The estimates of the coefficients for each variable are similar, and their 95% confidence intervals have a high coincidence. In my opinion, comparing the non informative piror and tradition OLS regression, the reason why we obtained similar conclusions is that

under the standard non information prior $p(\beta, \sigma^2) \propto 1/\sigma^2$, we can obtain that the posterior of coefficientβ and co-variance matrix V_β are both the same as the estimate of OLS regression.

Furthermore, Even when we use a hierarchical model to analyze the data, the structure of the dataset is easily captured by linear models, different types of regression models might perform similarly. This is particularly evident when there is a linear relationship between the features and the dependent variable. Therefore, OLS regression, Bayesian regression, and hierarchical models share certain properties or assumptions that lead them to perform similarly on certain datasets.

2 Background

Diabetes mellitus (DM), a prevalent non-communicable disease, represents a significant global health concern, necessitating the development of accurate early detection methods based on established risk factors[5]. The Diabetes dataset, a widely utilized benchmark dataset in regression analysis, was originally introduced in "Least Angle Regression" by *Efron et al.* in 2004, published in the *Annals of Statistics*[1]. This dataset comprises 442 instances and encompasses 10 baseline variables, including age, sex, body mass index (BMI), average blood pressure, and six blood serum measurements deemed pertinent to predicting diabetes progression after one year.

The primary objective of our project is to compare **Ordinary Least Squares (OLS)** regression and **Bayesian regression** techniques using this specific dataset. Our focus lies in identifying the most predictive variables associated with diabetes progression. Specifically, we aim to develop a model capable of predicting diabetes progression one year after the baseline measurement, quantified by a response of interest measure.

It is worth noting that the original literature does not provide information regarding the dataset's collection process, precluding discussions about potential biases introduced during sampling or suggestions for improving data collection methods. Nevertheless, a general best practice in data collection is to include a diverse range of samples to ensure the broad applicability of analytical conclusions. In our context, the absence of specific details about the data collection process does not impact our selection of prior distributions. Given the dataset's widespread applicability in regression problems, choosing it as our dataset remains a favorable and pragmatic choice.

2.1 Exploratory Data Analysis

A small part of the dataset is shown in Table 1. 442 diabetes patients were measured on age, sex, body mass index, average blood, pressure and six blood serum measurements including cholesterol, glucose, lamorigine, etc. The response of interest is a quantitative measure of the progression of diabetes one year after the

measurements were taken.

Patient	AGE	SEX	BMI	BP	S1(tc)	S2(ldl)	S3(hdl)	S4(tch)	S5(ltg)	S6(glu)	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.6	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	22.6	89	139	64.8	61	2	4.2	68	97

Table 1: Diabetes dataset sample. There are 442 patients in all with 10 baseline variables measured to predict the disease progression after one year represented quantitatively.

To start with, we check the distribution of each variable and pairwise correlation between these variables, and the results are shown in Figure 1. It can be observed that each variable follows a unimodal distribution, except for S4. Indeed, after closer inspection, S4 turns out to be the discretized ratio between S1 and S3, which explains why it seems to be a multimodel distribution from the plot. Indeed, we check the distribution of $\frac{S1}{S3}$, which follows unimodality.

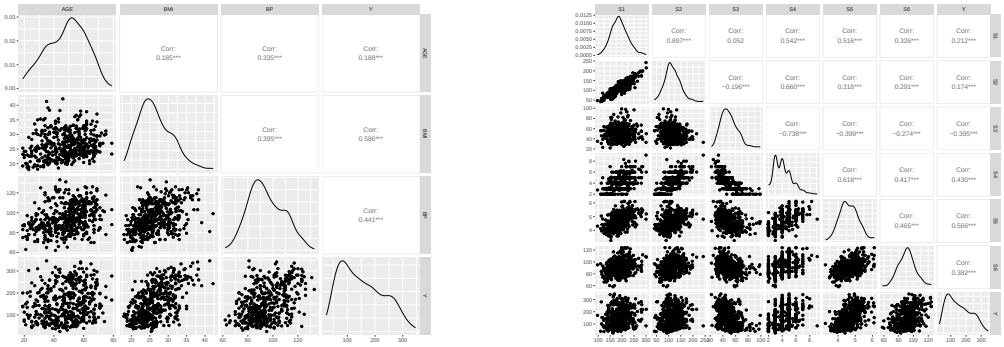

Figure 1: Exploratory analysis of the variables. The lower and upper corners demonstrate the scatter plots and the pairwise correlations between variables, respectively. The diagonal shows the estimated distribution of each variable.

Since sex is naturally a categorical variable, we also examine whether there is difference between male and female in Figure 2.

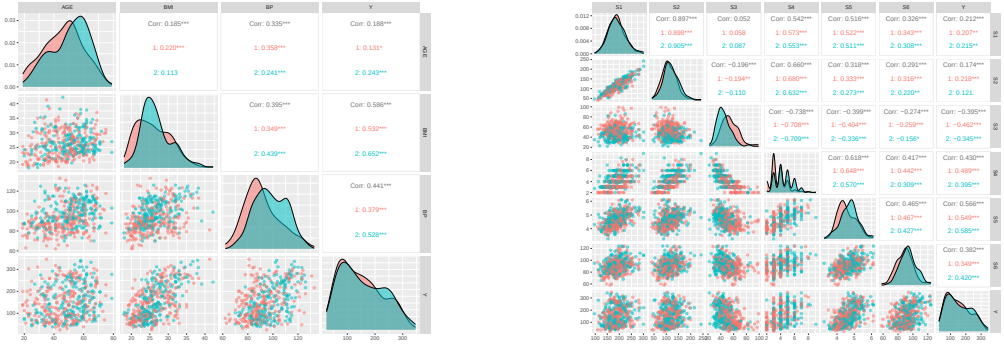

Figure 2: Exploratory analysis of the variables with sex as different groups. The layout is the same as Figure 1. Red represents male and green represents female.

Although there exists certain difference on body mass index and blood pressure, no obvious difference is observed among the blood serum measurements, which indicates there is little need to consider different gender separately.

3 Methods

3.1 Basic Settings

The analysis of our project is based on Bayesian linear regression model. Here, we list some basic settings.

In a linear regression model, we are interested in the relationship between X and y. Generally, we assume

$$X|\phi \sim p(X|\phi), y|X, \theta \sim p(y|X, \theta)$$

in which ϕ, θ are parameters, and they have no common components.

Specifically, for the likelihood, we adopt the assumptions of classical linear regression, which include linearity, independence, normality, and constant variance.

$$y|X, \theta \sim N(X\beta, \sigma^2 I).$$

Here we take $\theta = (\beta, \sigma^2)$. Moreover, we assume there is no error when measuring X.

For the prior of ϕ, θ, we assume

$$p(\phi, \theta) = p(\phi)p(\theta)$$

i.e. ϕ and θ are independent apriori.

Since we are more interested in the relationship between X and y, we focus on the parameter θ, From these assumptions, we have

$$p(\phi, \theta|X, y) = p(\phi|X)p(\theta|X, y)$$

So, to learn about θ, we only need to look at $p(\theta|X, y)$. In other words, our models can be conditional on X.

3.2 Choices of Prior

For the specific prior of θ, we plan to analyze the following choices [4]

3.2.1 Standard Noninformative Prior

$$p(\theta) = p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

3.2.2 Simple Conjugate Prior

$$\beta|\sigma^2 \sim N(m_0, \sigma^2 C_0)$$
$$\sigma^2 \sim \text{Inv} - \chi^2(v_0, s_0^2)$$

Here, σ^2 (the variance parameter) follows an inverse chi-squared distribution with v_0 degrees of freedom and scale parameter s_0^2. This is a common choice for a prior on a variance parameter in Bayesian analysis.

3.2.3 Hierarchical Conjugate Prior

$$\beta|\sigma^2 \sim N(m_0, \sigma^2 C_0)$$
$$\sigma^2 \sim \text{Inv} - \chi^2(v_0, s_0^2)$$

and

$$p(m_0, C_0) \sim \text{LKJ-dist}(\eta = 1)$$
$$v_0 \sim \text{Cauchy}(0, 1)\mathbb{I}(x > 0)$$
$$s_0^2 \sim \text{Cauchy}(0, 1)\mathbb{I}(x > 0)$$

This specifies a prior for the mean m_0 and covariance matrix C_0 using the LKJ distribution, which is often used for covariance matrices. The parameter $\eta = 1$ typically implies a uniform distribution over correlation matrices. Both v_0 and s_0^2 are distributed according to a Cauchy distribution centered at 0 with scale 1. The indicator function $\mathbb{I}(x > 0)$ suggests that only positive values are considered valid, which makes sense for parameters like degrees of freedom and variance.

The LKJ distribution allows for a wide range of correlation structures, which offers flexibility and control for modeling correlations, while the Cauchy distribution provides robustness and flexibility for scale parameters. These choices reflect a balance between informativeness and flexibility, allowing the data to play a significant role in informing the posterior distributions of the model parameters.

3.3 Model Selection

We plan to employ the model averaging method. To implement this, we introduce another indicator vector, denoted as Z. This vector shares the same length as the number of predictors in the dataset. We assign to it an i.i.d Bernoulli prior,

$$Z[i] \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(0.5)$$

Subsequently, we perform element-wise multiplication of this vector with β to obtain a new coefficient vector, denoted as β_{ind},

$$\beta_{\text{ind}}[i] = Z[i] \times \beta[i]$$

And our final model is determined by posterior mean of β_{ind}.

In this setting, we use the Hierarchical Conjugate Prior from above.

4 Results

In our project, we use OLS regression, bayesian regression and hierarchical model to analyze data and obtained the influence of each index on diabetes mellitus and the correlation between them.

4.1 OLS Regression Model

First of all, we have a look at all the data and do OLS regression and the results are shown in below table.

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($\geq \mid t \mid$)**
Intercept	152.133	2.576	59.061	<2
age	−10.012	59.749	−0.168	0.867 000
sex	−239.819	61.222	−3.917	0.000 104
bmi	519.840	66.534	7.813	4.30
map	324.390	65.422	4.958	1.02
tc	−792.184	416.684	−1.901	0.057 947
ldl	476.746	339.035	1.406	0.160 389
hdl	101.045	212.533	0.475	0.634 721
tch	177.064	161.476	1.097	0.273 456
ltg	751.279	171.902	4.370	1.56
glu	67.625	65.984	1.025	0.305 998
Residual Std. Error:	54.15 on 431 degrees of freedom			
Multiple R-squared:	0.5177			
Adjusted R-squared:	0.5066			
F-statistic:	46.27 on 10 and 431 DF, p-value: < 2.2e-16			

Table 2: Summary of OLS regression

From Table 2 we see that $R^2 = 0.5177$ is relatively moderate, indicating a reasonable fit but leaving room for improvement. I personally think that it is a not bad model but not good enough. Thus we use AIC as criteria of evaluation to choose appropriate variables and decide our final model.

From the stepwise regression process, we find that it removes age, hdl, tch and glu, and the final model with AIC criterion is

$$y = 152.1 - 226.5 \times sex + 529.9 \times bmi + 327.2 \times map - 757.9 \times tc + 538.6 \times ldl + 804.2 \times ltg \tag{1}$$

Df	Sum of Sq	RSS	AIC	
$< none >$	1271491	3534.3	-	
ldl	1	39378	1310869	3545.7
sex	1	41858	1313349	3546.6
tc	1	65237	1336728	3554.4
map	1	79627	1351119	3559.1
bmi	1	190586	1462077	3594.0
ltg	1	294094	1565585	3624.2

Table 3: AIC Stepwise Variable Selection Results

R^2 of the final model is 0.5149, and Adjusted R^2 is 0.5082. From the results we can see, it only improves a little comparing the original model, the adjusted R^2 increased by 0.0016, and the sum of squared residuals only decreased by 0.09.

4.2 Default Bayesian Regression

We employed the default Bayesian regression model to analyze this problem. As anticipated, and in accordance with theoretical expectations, this model yielded results remarkably similar to those obtained through OLS regression.

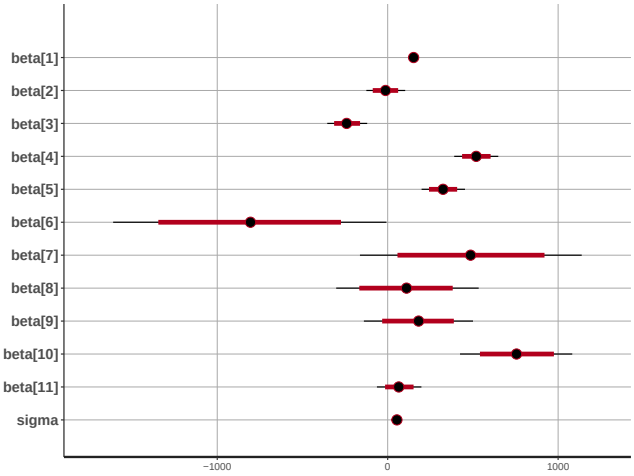

Figure 3: Confidence interval of explanatory variables for default Bayesian Regression.

4.3 Model with Simple Conjugate prior

In addition, we would like to introduce bayesian regression model and compare the results with OLS regression. As a consequence of the fact that the model is relatively simple and the data is limited, so it does not have much information about the parameters. Therefore, we choose a loose prior that does not have a

particular preference. This helps to avoid introducing too much information into the prior. the results are shown in Table 4.

Parameter	**Mean**	**se_mean**	**sd**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
Intercept	151.81	0.03	2.99	145.92	149.82	151.85	153.80	157.59
age	29.54	0.55	46.52	−61.07	−1.62	30.84	60.69	121.86
sex	−83.38	0.53	45.14	−173.08	−114.10	−83.68	−52.99	4.38
bmi	306.85	0.61	46.13	216.25	275.93	306.41	337.45	398.21
map	201.36	0.59	47.07	111.46	169.22	201.12	232.92	294.82
tc	6.03	0.73	52.47	−97.92	−28.36	7.37	41.57	105.90
ldl	−29.59	0.69	51.56	−130.39	−63.98	−29.94	5.97	71.78
hdl	−150.49	0.65	49.64	−249.23	−184.27	−150.44	−117.94	−52.64
tch	118.21	0.75	52.50	15.57	82.38	117.47	154.11	219.27
ltg	263.47	0.62	49.58	167.43	230.51	263.24	295.60	360.79
glu	111.78	0.55	46.07	21.79	81.52	111.75	143.32	200.59
sigma2	3898.91	2.99	267.41	3414.87	3712.84	3886.06	4067.51	4483.70

Table 4: Bayesian Regression Results

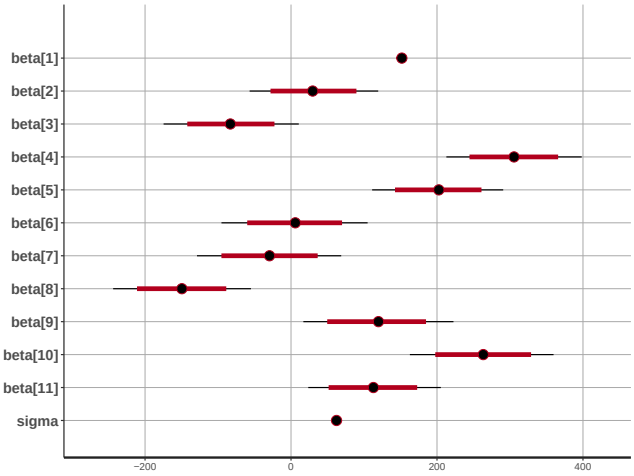

Figure 4: Confidence interval of explanatory variables for Bayesian regression model.

4.4 Model with Hierarchical Conjugate prior

In Bayesian linear regression, we see that the results we get are similar to the traditional linear regression results, and we do not get better results, so next we use the Bayesian hierarchical model. Bayesian hierarchical models provide a richer framework for inference. Sampling from probability distributions allows us to obtain the entire distribution of parameters, not just point estimates. This aids in understanding the uncertainty associated with parameters and allows for more comprehensive inference. Then we choose LKJ Distribution for Covariance Matrices $p(m_0, C_0)$. It is specifically designed for modeling correlation matrices. It allows for a wide range of correlation structures, making it a versatile choice for various applications. Also, we choose cauchy distribution for scale parameters v_0, s_0^2 because of Heavy Tails: The Cauchy distribution is known for its heavy tails compared to the normal distribution. This characteristic is useful when the parameter might have extreme values, as the heavy tails allow for a wider range of plausible values. It's especially suitable for scale parameters like variance, where we often expect a broad range of plausible values. Finally, we construct a hierarchical model to analyze data. The results are shown in Table5

Parameter	**Mean**	**sd**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
Intercept	152.17	2.65	146.93	150.45	152.16	153.86	157.56
age	−7.58	58.38	−124.36	−45.95	−6.33	30.61	108.39
sex	−239.65	61.97	−357.80	−279.50	−239.54	−198.92	−116.05
bmi	520.24	67.35	388.56	473.82	520.48	566.01	650.33
map	323.58	66.05	191.74	279.24	322.31	368.98	454.14
tc	−802.63	427.73	−1654.91	−1088.66	−807.79	−515.97	45.71
ldl	485.52	349.04	−196.39	253.62	486.05	712.62	1201.15
hdl	103.65	217.21	−330.04	−39.04	105.76	250.26	510.56
tch	175.68	161.11	−148.58	71.32	176.17	286.40	493.45
ltg	756.16	173.59	413.44	640.63	754.82	874.80	1089.01
glu	66.64	66.77	−59.80	21.36	66.90	112.38	195.22
sigma2	2943.88	203.71	2573.60	2801.74	2931.99	3070.55	3386.29

Table 5: Bayesian Regression Results with Hierarchical Model

We were intrigued to discover that our results closely align with those obtained from the default Bayesian regression model. This similarity, we theorize, stems from our use of non-informative hyperpriors. By mitigating the impact of prior information on the parameters, these hyperpriors ensure that our results are more heavily influenced by the data itself. Hence, it's logical to conclude that the data plays a more significant role in shaping our findings.

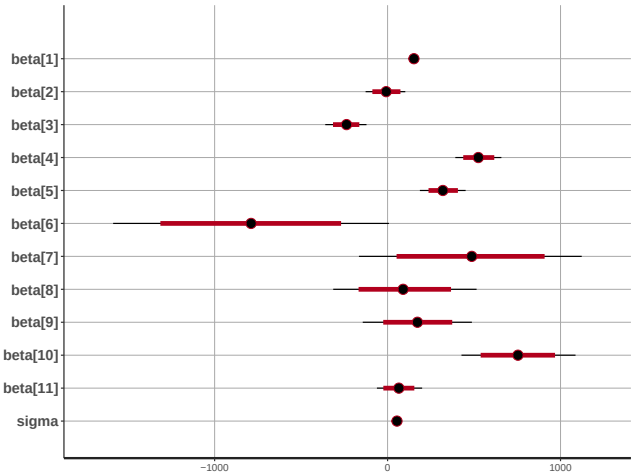

Figure 5: Confidence interval of explanatory variables for hierarchical model.

5 Conclusion

In this study, we applied different regression models to understand the influence of various factors on diabetes mellitus. Through our comprehensive analysis using Ordinary Least Squares (OLS) regression, Bayesian regression, and a Bayesian hierarchical model, we gained valuable insights into the relationships between these factors and the disease.

The OLS regression model provided a baseline for our analysis. Although the model showed a moderate fit with an R^2 of 0.5177, it indicated that improvements could be made. By applying the AIC criteria for variable selection, we refined the model, achieving a slight increase in the Adjusted R^2 and a marginal decrease in the sum of squared residuals. However, the improvement of this model compared to the original model is still very limited, so we hope to try other methods to get a better model.

The Bayesian regression model introduced a probabilistic approach, allowing for a more nuanced understanding of parameter uncertainties. However, the results were similar to those of the OLS model, suggesting that the additional complexity of the Bayesian approach did not significantly alter the insights gained.

The Bayesian hierarchical model provided the most sophisticated analysis. By incorporating hierarchical structures and broader distributional assumptions, we gained a richer and potentially more accurate understanding of the data. This model offered a more flexible approach to understanding the complex interactions between variables.

Our findings suggest that while traditional models like OLS can provide valuable insights, advanced models like Bayesian hierarchical models can offer different understanding. In our research, though we use different methods to analyze data, we get similar results. I think that the marginal improvements observed with these advanced techniques also indicate that the simplest model might often be sufficient for practical purposes.

References

[1] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of Statistics 32*, 2 (2004), 407 – 499.

[2] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*, 2nd ed. ed. Chapman and Hall/CRC, 2004.

[3] Hoff, P. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer New York, 2009.

[4] Kruschke, J. K. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science 1*, 5 (2010), 658–676.

[5] Olokoba, A. B., Obateru, O. A., and Olokoba, L. B. Type 2 diabetes mellitus: a review of current trends. *Oman Med J 27*, 4 (July 2012), 269–273.