Evaluation of Bayesian and Frequentist Methods for Vitamin D Deficiency Prediction

Sana Gupta

sana.gupta@uconn.edu

BIST 5615

December 20, 2024

Abstract

Vitamin D deficiency is prevalent in a large number of people and can cause serious health problems, including rickets, depression, hair loss, and other health issues. In this paper, we evaluate the strength of the relationship between various risk factors and Vitamin D deficiency. We use a frequentist ordinal logistic regression model and a Bayesian logistic model with priors based on existing literature. Both models identified taking Vitamin D supplements and race as having a significant impact on odds of having a Vitamin D deficiency. Furthermore, the frequentist model identified that taking diet medication has a positive relationship with an increased chance of being deficient. The Bayesian model identified non-daily milk consumption and not being college educated as significant positive predictors of having a Vitamin D deficiency. Both models were evaluated and had similar model fit metrics, with the Bayesian model falling slightly behind the frequentist model.

1 Introduction

1.1 Vitamin D

Vitamin D is a naturally-occurring vitamin that can be found a few foods or taken as a supplement. Ultraviolet rays from sunlight can also cause Vitamin D synthesis when they hit human skin. Vitamin D is important for bone growth and general bone health, and not having enough of it can result in brittle bones and conditions such as rickets and osteoporosis. A lack of Vitamin D is also known to cause depression [1].

There are many factors that can cause Vitamin D deficiency. For example, not consuming enough calcium through milk or other calcium-rich foods can lead to a deficiency. Additionally, those with dark skin tones have trouble absorbing Vitamin D through UV rays and are often more likely to have a deficiency than those with lighter skin. The same can be said about those who live in climates with less sunlight, or people who don't go outside often.

Treatment for Vitamin D deficiencies is often through Vitamin D supplements. Lifestyle changes such as spending more time in the sun without sunscreen (when the UV index is low) can also help improve Vitamin D levels. However, before doctors can recommend this treatment, the deficiency needs to be diagnosed. With so many risk factors, it is likely that many people have a Vitamin D deficiency and are not aware about it. The goal of this paper is to propose two different models that could be used as clinical predictive tools. These models could be used to inform patients and their doctors about their risk level for Vitamin D deficiency based on their lifestyle and demographics. Those with higher risk of being deficiency could then go get tested and begin treatment [1].

1.2 Types of Models

The logistic regression is a statistical tool that can be used to model the relationship between a categorical dependent variable and a number of independent variables, which can be categorical or numeric. By using the logistic function, these models are able to make sure that any predicted probabilities are between 0 and 1, the bounds for probabilities. Lead-in to the Two Types

Standard binary logistic regressions are common in a wide variety of fields, but they are used for binary outcomes, which means the outcome has only two potential values

(Yes/No, Survival/Death, etc). Ordinal logistic regressions are an extension of binary logistic regressions, where the outcome variable can have more than one level, and the levels are naturally ordered. We propose the use of a frequentist ordinal regression and a Bayesian ordinal logistic regression model for this analysis.

Ordinal logistic regressions are based off of the frequentist framework of the binary logistic regression. The coefficients produced by the model can be interpeted as the log-odds of being at one level of the response category versus being in a higher category. The fences between each of the response levels are called thresholds.

Bayesian ordinal logistic regressions take this idea a step further by allowing prior information or outside research to be incorporated. Using odds ratios from prior results or any other kind of domain knowledge, we can assign prior distributions on any of the model parameters. In doing so, the parameters are weighted by the importance of the priors in relation to each other.

2 Motivating Case Study

My dataset is from the 2015-2016 National Health and Nutrition Examination Survey. This annual survey collects information about the demographics, health, and nutritional habits of American people. The survey involves face-to-face interviews as well as physical examinations and lab results in order to fully understand the ever-changing health climate in America. The data can also be used to assess trends in health across different demographic groups or other factors.

Through my data cleaning and organizing process, I compiled information on various risk factors that could explain Vitamin D deficiency. In total, my dataset had 3277 observations. The Vitamin D lab measurement was used for the response variable. In the dataset, this was reported numerically and I categorized the values using outside research on cuttoffs for deficiency levels as follows:

- Sufficient: Measurement ¿ 50 nanomoles per liter
- Mild Deficiency: 30 ¡ Measurement ¡ 50 nanomoles per liter
- Severe Deficiency: Measurement ¡ 30 nanomoles per liter

The response variables included in the dataset include:

- Taking Vitamin D Supplement
- Taking Calcium Supplement
- Taking Weight Loss Pills
- Having Weight Loss Surgery
- Consumption of Vit. D in Diet (Day 1 and Day 2)
- Race
- Milk Consumption
- Education Level

Figure 1: Bar chart showing the distribution of Vitamin D deficiency levels.

Figure 1 shows the distribution of the three levels (sufficiency, mild deficiency, severe deficiency) in the dataset. It is worth noting that there is an imbalance in the levels. There are very few people in the severe deficiency group, and most people fall in the mild deficiency group.

3 Methods

In order to investigate which variables can be best used to predict Vitamin D deficiency level, a logistic regression can be used. Specifically, I chose to use ordinal logistic regressions because of the ordered nature of the response variable. I will compare the significant covariates and overall model performance of the Frequentist and Bayesian models.

3.1 Frequentist Approach

The Frequentist approach is based on an ordinal regression model where the response variable, Vitamin D deficiency, has three levels and is regressed against the levels of all the other covariates. After generating the first version of the model, the stepAIC() function in R was used to reduce the model to only the most significant covariates. This function works by eliminating variables one at a time in order to reduce the overall model AIC.

3.2 Bayesian Approach

The Bayesian model is similar to the Frequentist model, with the addition of informative priors with the goal of adding extra weight to the covariates that are known to be impactful for predicting Vitamin D deficiency. I based my priors on a paper titled "Prevalence and correlates of vitamin D deficiency in US adults". In this paper, the following covariates were identified as having a significant impact on the odds of having a Vitamin D deficiency, along with the associated odds ratios [2]:

- Identifying as Black (OR = 9.6)
- Identifying as Hispanic (OR = 3.2)
- No college education (OR = 1.3)
- Not consuming milk products daily (OR = 1.6)

I used the "brms" package in R to conduct the Bayesian regression analysis. The model set-up was similar to the frequentist model with the addition of normal priors on the above four covariates. For the means of the normal priors, I used the log of the odds ratio given in the paper, and I set the variance for the priors at 0.1 for the two

race categories and milk consumption, and 0.2 for the non-college educated group. This difference in the variance was based on the significance of the odds ratios given in the paper, and the three odds ratios that had p-values less than 0.001 were assigned a smaller variance in order to make the impact of the prior stronger [2].

After running the model with these priors, I manually selected only the covariates that had a significant impact on the model. These covariates make up the final reduced model.

4 Results

The Frequentist reduced model coefficients show that taking Vitamin D supplements, and identifying as Hispanic, White, Asian, or Multi-Racial decreases an individual's odds for having a Vitamin D deficiency. On the other hand, the model identified that people who identify as Black or take Vitamin D levels have an increased chance of having a Vitamin D deficiency.

Variable	**Coefficient**	**95% CI**
Sufficiency (Mild Deficiency)	-2.106	[-2.324, -1.889]
Mild Deficiency (Severe Deficiency)	2.004	[1.779, 2.229]
taking_vitd (Yes)	-1.638	[-1.798, -1.480]
race2 (Hispanic)	-0.236	[-0.517, 0.044]
race3 (White)	-1.412	[-1.641, -1.186]
race4 (Black)	0.154	[-0.111, 0.420]
race6 (Asian)	-0.075	[-0.363, 0.286]
race7 (Multi-Racial)	-0.698	[-1.123, -0.277]
took_diet_meds (Yes)	0.658	[0.059, 1.261]

Table 1: Frequentist Reduced Model Coefficients

The Bayesian model coefficients show that taking Vitamin D supplements, and identifying as White or Multi-Racial decreases an individual's odds of having a Vitamin D deficiency. Identifying as Hispanic, Black, or Asian, not consuming milk daily, and not being college educated are associated with a higher risk of Vitamin D deficiency.

The models were evaluated using two different metrics. First, the Log-Likelihood was

Variable	**Coefficient**	**95% CI**
b_Intercept[1]	-1.562	[-1.818, -1.310]
b_Intercept[2]	2.594	[2.333, 2.858]
b_taking_vitd (Yes)	-1.655	[-1.816, -1.499]
b_race2 (Hispanic)	0.448	[0.226, 0.669]
b_race3 (White)	-0.930	[-1.137, -0.727]
b_race4 (Black)	0.951	[0.735, 1.165]
b_race6 (Asian)	0.447	[0.137, 0.749]
b_race7 (Multi-Racial)	-0.197	[-0.610, 0.217]
b_milk_consumption (Not Daily)	0.185	[-0.126, 0.178]
b_education (Not College Educated)	0.103	[-0.071, 0.276]

Table 2: Bayesian Reduced Model Coefficients

computed for both methods. The goal of this metric is to measure the goodness-of-fit of the model for the observed data. Values that are less negative indicate a better fit. The Frequentist model has a Log-Likelihood of -2342.537, and the Bayesian model has a Log-Likelihood of -2368.154. These values are very close, indicating that there is not a significant difference between the goodness-of-fit for the two models.

The Akaike Information Criterion (AIC) is a Frequentist metric used to evaluate model selection, and it can be compared to the Bayesian Leave-One-Out Information Criterion (LOO-IC). The AIC for the Frequentist method is 4703.074, and the LOO-IC for the Bayesian method is 4746.061. Again, these values are not significantly different, indicating that the coefficients in both models are selected similarly well.

Figure 2 shows side-by-side Reciever Operating Characteristic (ROC) curves for each of the levels of Vitamin D deficiency for both models. Each curve is associated with an area under the curve (AUC) measurement, which is given in the legend next to each level that is being evaluated. AUC values close to 1 indicate optimal ability to discriminate between the correct and incorrect predictions for each observation. Computing AUC and making an ROC curve is straightforward for Frequentist models. For the Bayesian model, the Bayesian ROC curves were calculated by extracting the posterior predicted probabilities for each category from the Bayesian model, and then using the mean probabilities across posterior samples as the predicted probabilities for each observation. These were

compared to the true class labels to generate the ROC curves and compute AUC values. As seen in the figure, the models both perform similarly for all three levels, but both models struggle to predict Mild Deficiencies correctly.

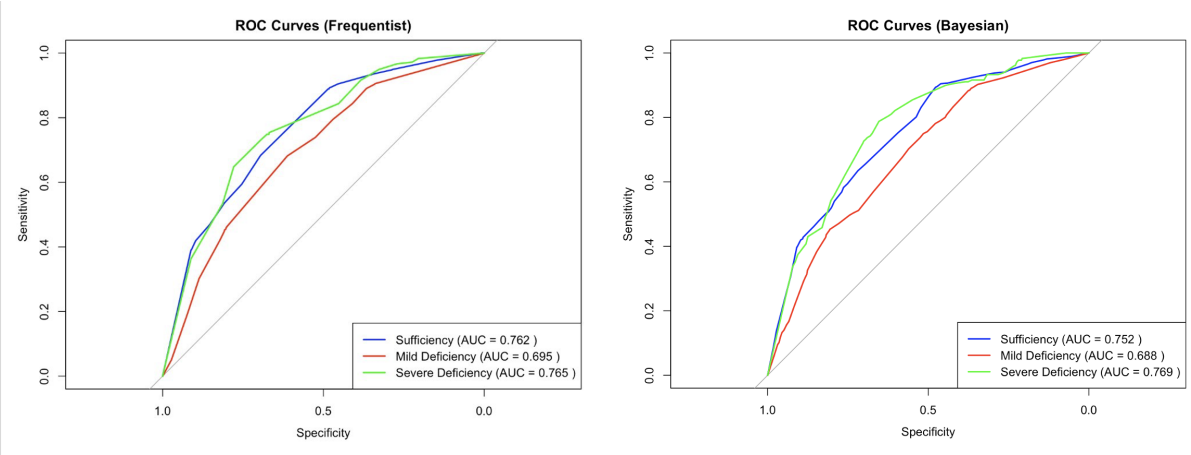

Figure 2: ROC Curves comparing the performance of the two models in differentiating between the three levels of Vitamin D deficiency.

5 Discussion

In conclusion, we have proposed two models that could be used as diagnostic tools to predict Vitamin D deficiency. One model is a Frequentist ordinal regression, and the other is a Bayesian ordinal regression with priors on some of the covariates based on outside research. The models were then reduced to only include significant parameters. There is some overlap between the significant parameters in both models, while there are other parameters that only show up in one model or the other. The performance and fit of the models is similar, with the Frequentist model slightly outperforming the Bayesian model in Log-Likelihood and AIC/LOO-IC metrics.

Because the Bayesian model includes the use of priors based on work from an outside publication, the coefficients associated with those priors ended up being significant enough to be a part of the final reduced model. While these priors make sense logically (ex. it makes sense that not drinking enough milk causes lower Vitamin D), it is possible that broader research on this topic and incorporating information from more than one source could allow for a more robust set of priors. Further research into the shape and strength

of the priors could also further improve the Bayesian model.

By comparing numerical metrics and the ROC curves in Figure 2, it is clear that the Frequentist model and Bayesian model both perform very similarly despite having a different set of parameters. It is worth noting that both models include taking Vitamin D supplements as being associated with decreased odds of having a Vitamin D deficiency, indicating that this parameter could be so significant that it outweighs the other parameters that are different in the two models. There is also some overlap in particular ethnicity groups, which could lead to a similar conclusion that these few predictors are so significant that the inclusion of other predictors in either model is just not that important.

From the ROC curves, we also see that both models struggle to identify participants who only have a Mild Deficiency. From Figure 1, we can visualize the issue of unbalanced classes. There are significantly fewer individuals in the dataset who have a Severe Deficiency than either of the other two groups, and it is known that Bayesian ordinal models perform better when there is more balance in the dataset. Future work could involve adding weighting or using other methods to balance the dataset.

Despite these potential issues, these models could serve as the foundation for important diagnostic tools. There are a lot of people who have Vitamin D deficiencies, but so many individuals are not aware of their deficiency because they have never been formally tested by a doctor. A diagnostic tool that predicts the odds of having a Vitamin D deficiency could motivate individuals to go get tested if they are flagged by the models as being at high risk. Furthermore, a tool like this could be more accessible to people than regular consultations with a doctor, allowing them to take control of their health without needing to deal with barriers to accessing healthcare.

Acknowledgments

Thank you to Professor Ming-Hui Chen for teaching us this semester and giving feedback to help improve our projects!

References

[1] Screening for vitamin d deficiency: A systematic review for the u.s. preventive services task force. *Annals of Internal Medicine*, 162(2):109–122, 2015. PMID: 25419719.

[2] Kimberly Y.Z. Forrest and Wendy L. Stuhldreher. Prevalence and correlates of vitamin d deficiency in us adults. *Nutrition Research*, 31(1):48–54, 2011.