Investigating Association between BMI and Other Variables Using Skew-Symmetric Regression Models

Kaolee Yang | University of Wisconsin – Eau Claire

Introduction

Finding significant predictors of body mass index (BMI) drew researcher attention for decades due to the well-known fact that very high BMI leads to obesity [1]. Most of the previous researcher use ordinary least squares (OLS) regression for this purpose in which the error term follows normal distribution [2]. Since the distribution of BMI is skewed [3], ordinary regression may not be suitable to determine significant covariates of BMI. In this project we used two real life datasets and used multiple skew-symmetric regression model to identify variables that affects BMI. We also compared our results obtained from the skew-symmetric models to the results from OLS regression.

Methods

Skewed Normal Distribution

Skew-normal distribution was introduced by Azzalini [4]. If Y follows skew normal distribution then probability density function of Y is

$$f(x) = 2\varphi(x)\Phi(\alpha x), x \in R \qquad (1)$$

Where α is the shape parameter and is a real number. The skew normal densities for varying shape parameters are plotted in Figure 1. As observed from the figure, normal distribution is a special case when $\alpha = 0$, and for other selected values of alpha, either a positive or negative skewness is observed.

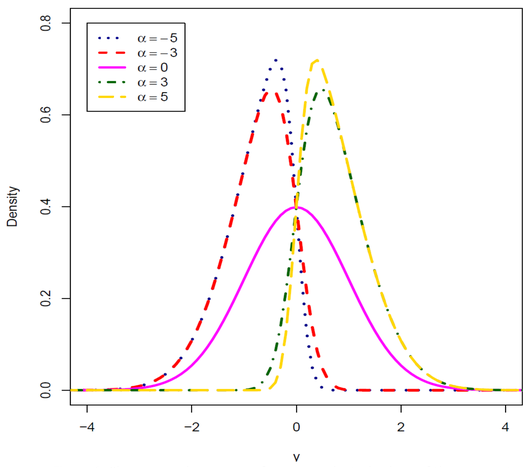

Figure 1: Skew normal regression for some selected values of the parameter α

Skew-Symmetric Regression and OLS Regression

The multiple skew-symmetric regression model is given by

$$Y_j = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ij} + \alpha Z_i + \epsilon_j \qquad (2)$$

with $\epsilon_j \sim N(0, \sigma^2)$ and $z_j \sim HN(0,1), j = 1, \cdots, n$ all independent, where $HN(0,1)$ denotes univariate standardized half normal distribution. Note that $\alpha Z_i + \epsilon_j \sim SN(0, \sigma^2, \alpha)$ and $Y_j \sim SN(X_j^T \beta, \sigma^2, \alpha)$ with $X_j^T = (1, X_{j1}, \cdots, X_{jp})$ and $\beta = (\beta_0, \beta_1, \cdots \beta_p)^T$.

The multiple normal regression (OLS) model is obtained from (2) taking $\alpha = 0$.

Applications

Data Description

To determine the significant predictors related to BMI we examined the following two data sets namely UWEC BMI data and NHANES data.

The UWEC BMI data was retrieved from the National College Health Assessment conducted by the UWEC Student Health Service. The data we analyze comes from a sample of 630 students attending UWEC during the 2014-2015 academic year and 6 related variables.

The other BMI data set comes from the National Health and Nutrition Examination Survey (NHANES), collected by the US National Center for Health Statistics (NCHS) and is available in the R package nhanes. The data were taken from the 2009-2010 and 2011-2012 sample years. The original data set contained 10,000 rows of observed BMI records and 75 variables. However, to determine the relationship between BMI and other variables we generated 5000 random rows of data and considered 17 variables in our sample. After deleting the missing values, we use $n = 441$ for our final analysis.

Descriptive Statistics

The descriptive statistics for UWEC BMI and NHANES BMI data are shown in Table 1 and 2.

Table 1: Summary Statistics for UWEC BMI Data

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
17.47	21.46	23.40	24.65	26.32	52.87

Table 2: Summary Statistics for NHANES BMI Data

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
15.02	23.60	26.96	28.08	31.38	59.10

Respectively, we also plotted histograms to get initial idea of the BMI data. Figure 2 represents the histograms. The histograms clearly shows that both the data are positively skewed (right-skewed).

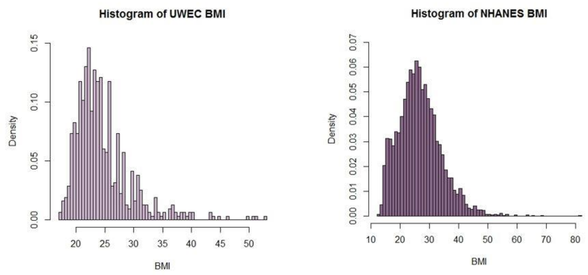

Figure 2: Histograms of UWEC BMI and NHANES BMI

In addition, we constructed normal probability plots for both data that are shown in Figure 3. The normal probability plots reveal that the assumption of normality for OLS regression may not be satisfied . Shapiro-Wilk normality tests with P-values of < 0.0001 and < 0.0001 respectively confirm our observations.

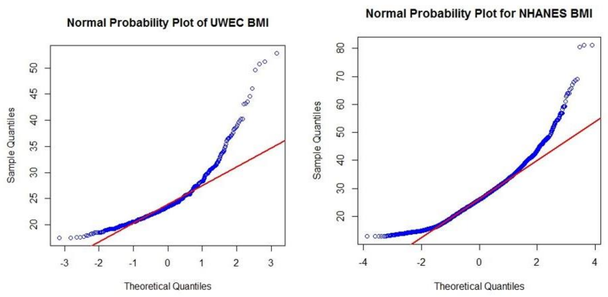

Figure 3: Normal probability plots of UWEC BMI and NHANES BMI

Results: UWEC Data

Of the variables considered for UWEC BMI, we found that the only significant variables in determining BMI distribution were height *($p < 0.0001$)* and weight *($p < 0.0001$)*. This does not come as a surprise because BMI is defined as the ratio of body weight and body height. The rest of the covariates for UWEC BMI were found insignificant.

Table 3: P-values from skew-symmetric and OLS | UWEC BMI Data

Factors	P-Value (SN)	P-Value (OLS)
Height	< 0.0001 ***	< 2e-16 ***
Weight	< 0.0001 ***	< 2e-16 ***
Age	0.5442	0.339
Alcohol	0.1912	0.273
Gender (Male)	0.4481	0.552
Gender (Other)	0.9892	0.989
Physical Activity	0.8043	0.264

Note: *** $p = < 0.001$, ** $p = < 0.01$, * $p = < 0.05$

Comparing the p-values from the skew-symmetric models to the OLS regression, we found that both models identified the same variables as being significant. However, the Akaike information criterion (AIC) value (1047.670) for skew-symmetric model clearly shows a better fit than OLS model (AIC = 1089.883).

Results: NHANES Data

Of the variables considered for NHANES BMI, we found that height *($p < 0.0001$)* and weight *($p < 0.0001$)* illustrated the most significant relationship with BMI.

Alongside these two variables, SleepTrouble *($p = 0.0110$)*, Poverty *($p = 0.0041$)*, Education 9-11thGrade *($p = 0.0034$)*, Education$HighSchool *($p = 0.0082$)*, Education$SomeCollege *($p = 0.0178$)*, Race1$Hispanic *($p = 0.0070$)*, and Race1$Other *($p = 0.0021$)* were also highly significant predictors of BMI.

The other variables found significantly related to BMI were, SmokeNow *($p = 0.0217$)*, SleepHrsNight *($p = 0.0500$)*, Race1$Mexican *($p = 0.0411$)*, and Race1$White *($p = 0.0145$)*.

Table 4: P-values from skew-symmetric and OLS | NHANES BMI Data

Factors	P-Value (SN)	P-Value (OLS)
HardDrugs	.50890	0.37479
SmokeNow	.02170 *	0.01497 *
AlcoholDay	0.50802	0.63718
PhysActiveDays	0.43818	0.53642
PhysActive	0.41245	0.43017
SleepTrouble	0.01104 **	0.00764 **
SleepHrsNight	0.05003 *	0.04552 *
Diabetes	0.79018	0.84686
Pulse	0.54696	0.75871
Height	< 0.001 ***	< 2e-16 ***
Weight	< 0.001 ***	< 2e-16 ***
Poverty	0.00412 **	0.00238 **
Marital Status (LivePartner)	0.83709	0.83737
Marital Status (Married)	0.91176	0.97118
Marital Status (NeverMarried)	0.52338	0.64647
Marital Status (Separated)	0.88289	0.90626
Marital Status (Widowed)	0.72602	0.73020
Education (9 – 11th grade)	0.00344 **	0.00658 **
Education (CollegeGrad)	0.12350	0.16158
Education (High School)	0.00820 **	0.01011 *
Education (SomeCollege)	0.01782 **	0.01816 *
Race1 (Hispanic)	0.00696 **	0.00838 **
Race1 (Mexican)	0.04106 *	0.05163
Race1 (Other)	0.00212 **	0.00320 **
Race1 (White)	0.01450 *	0.02114 *
Age	0.11440	0.08399
Gender (Male)	0.25227	0.51563

Note: *** $p = < 0.001$, ** $p = < 0.01$, * $p = < 0.05$

Also comparing the p-values from the skew-symmetric models to the OLS regression for NHANES data, we found that skew-symmetric distributions best identifies significant variables that affects BMI. While OLS is competitive in determining significant factors that influence the shape and modality of BMI data, skew-symmetric models are more suitable. Again, AIC value of 799.1723 shows a better fit for the skew-symmetric model compared to OLS model (AIC=862.0908).

Conclusion

The goal of our study was to model body mass index (BMI) for a given set of covariates using skew-symmetric regression models in order to determine the factors that affect BMI. We found that height and weight were the most significant variables that affect BMI. Besides height and weight, some other variables that shows significant relationship with BMI were sleep troubles, sleep hours per night, race, education, smoking, poverty status, etc.

Furthermore, our comparison of skew-symmetric models to the OLS regression indicates that for datasets that exhibits asymmetry, the skew-symmetric models are more suitable. Our findings confirm that skew-symmetric models were better able to identify significant variables at varying levels. Although the OLS may not be suitable for skewed data such as BMI, it still appears to be a good tool for finding significant predictors of BMI.

We are hopeful that the framework presented in this paper for testing the significance of a related set of covariates with BMI will provide us better understanding of the distribution of BMI in terms of factors that influence it. Finally, we had to delete lots of missing values for NHANES data. For future research, some other skewed models can be chosen that can incorporate the missing values for more accurate results.

Acknowledgements

We thank the Office of Research and Sponsored Programs and the UWEC Mathematics Department for their support of our project. We also thank Student Support Services at UWEC for providing the data. Additionally, we thank Learning and Technology Services for printing this poster.

References

[1] Jackson, A. S., Stanforth, P. R., Gagnon, J., Rankinen, T., Leon, A. S., Rao, D. C., . . . Wilmore, J. H. (2002). The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *International Journal of Obesity*, 26, 789-796. doi:10.1038/sj.ijo.0802006

[2] Bottone, F. G., Jr., Hawkins, K., Musich, S., Cheng, Y., Ozminkowski, R. J., Migliori, R. J., & Yeh, C. S. (2013). The relationship between body mass index and quality of life in community-living older adults living in the United States. *J Nutr Health Aging*,17(6). doi:10.1007/s12603-013-0022-y

[3] Tran, T., Wiskow, C., and Aziz, M. (2017). Skewed and Flexible Skewed Distributions: A modern look at the distribution of BMI. *American Journal of Undergraduate Research (*to appear in June/July issue)

[4] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12, 171-178.

Advisor: Dr. Mohammad Aziz, Department of Mathematics, University of Wisconsin – Eau Claire