Bayesian Skew-Normal and Skew-t Models of Birth Weight and Food Security

Carter Allen; Brian Neelon, PhD; Sara E. Benjamin Neelon, PhD, JD, MPH

Department of Public Health Sciences, Medical University of South Carolina

Objectives

We examine the properties of skew-normal and skew-t models from both a Bayesian and frequentist perspective, and investigate the computational tools available for fitting these models. We apply skew-normal and skew-t models to data from the Nurture study, a cohort of mothers who gave birth between 2013 and 2016, where we seek to model the effect of food security during pregnancy on birth weight.

Introduction

In many applications of classical linear regression, the distribution of residuals exibits non- normal qualities such as skewness or heavy tails, making the assumption of normal error terms difficult to justify. The common statistical suggestion in these cases is to implement a transformation of the response variable, but this can result in a loss of interpretability. The skew-elliptical family is a broad class of probability distributions that contain the normal distribution as a special case and allow for flexible modeling when data exhibit skewness.

Definitions

Let ϕ and Φ be the standard normal pdf and cdf, respectively. Azzalini (1985) defined the density of a skew-normal random variable Z follows.

$$f(z; \lambda) = 2\phi(z)\Phi(\lambda z)$$

Similar to the construction of the familiar student's t random variable, we (cite) can define a skew-t random variable as the ratio of a skew normal and the square root of a χ^2 divided by its degrees of freedom. The resulant density is

$$t(x; \lambda, \nu) = 2t_0(x; \nu)T_0(\lambda x\sqrt{\frac{\nu+1}{\nu+x^2}}; \nu+1)$$

where t_0 and T_0 are the density and mass functions of the student's t distribution, respectively. A linear regression model with skew error terms is a modification of classical regression with the modification of either assuming $\mathcal{SN}$ or $\mathcal{ST}$ random errors.

Motivation

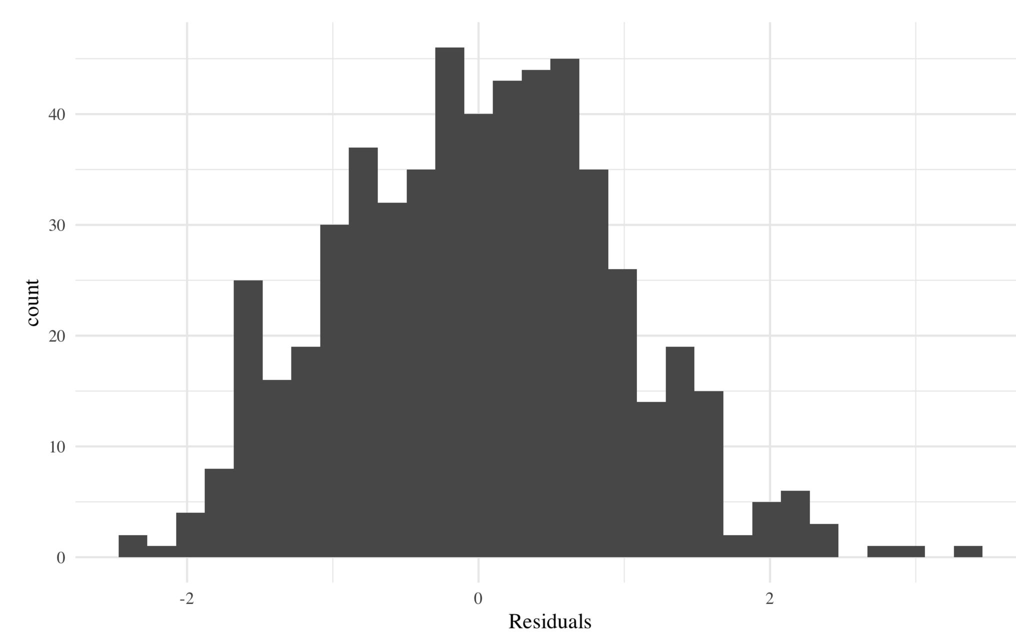

Figure 1: Distribution of residuals exhibiting skewness. Pearson's skewness = 0.17. Shapiro-Wilk p-value = 0.08.

Comparison Criteria

Important Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Modeling Approaches

1. *Maximum Likelihood Estimation*: For a random sample $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} \mathcal{SN}(\xi, \omega^2, \alpha)$, where $\xi = \mathbf{x}^T\beta$ for a collection of predictors $x_1, x_2, ..., x_n$ and a vector of unknown regression coefficients $\beta \in \mathbb{R}^p$. Azzalini (2014) describes procedures for obtaining MLE estimates of β and α, our primary parameters of interest. Azzalini's `R` package `sn` contains a function `selm` for fitting regression models with $\mathcal{SN}$ or $\mathcal{ST}$ random errors.
2. *Bayesian Gibbs Sampler*: We introduce the following stochastic representation of the skew normal distribution

$$Y_i = \mathbf{x}_i\beta + \psi z_i + \sigma\epsilon$$

Imputation Results

truncSamp.png

Figure 2: Simulated model parameter estimates method 2

Method	Abs. Err.	Bias	Variance
1	8.417	0.576	0.229
2	2.187	-0.009	0.002
3	9.476	-0.493	1.1619

Modeling Results

4.png

Figure 3: Model summary

3.png

Figure 4: Residual analysis

Conclusion

Missing and censored values are best imputed by using Method 2, *Sampling from a Truncated Normal*. A general lineral model can be built after imputing according to Method 2, and used in the future to predict *Enterococcus* concentrations in Charleston, SC.

Further Resources

https://carter-allen.github.io