A Comparison of Skew-Normal and Skew-t Regression Models

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.

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.

Motivation

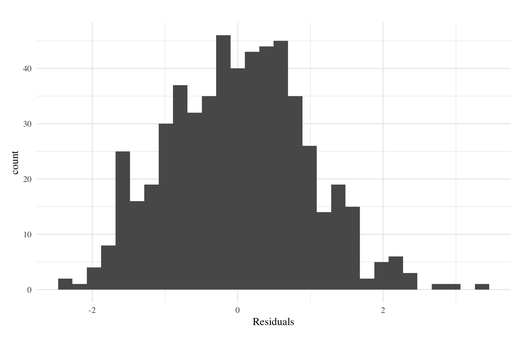

Figure 1: Distribution of residuals exhibiting skewness. Pearson's skewness = 0.17. Shapiro-Wilk p-value = 0.08.

Important Results

Enterococcus counts are significantly influenced by precipication, tidal stage, and location. The best performing imputation method of the three tested employs sampling from a truncated normal distribution.

Imputation Methods

1. *A naive approach*: Censored values are sampled from a $Uniform(0, 10)$ and missing values are replaced with the overall mean *Enterococcus* for 2013-2015.
2. *Sampling from a Truncated Normal*: Censored values are sampled from a $truncNorm()$ distribution, and missing values are predicted by the model made from all non-missing values.
3. *Expected value of a Truncated Normal*: Let $Y =$ *Enterococcus* concentrations. If $Y \sim logNormal(\mu, \sigma^2) \implies ln(Y) \sim Normal(\mu, \sigma^2)$, it can be shown that $E[ln(Y)|a < ln(Y) < b] = \mu + \sigma\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$. When $a = -\infty$ and $b = ln(10)$, $E(ln(Y)|0 < ln(Y) < ln(10)) = \mu - \sigma\frac{\phi(\frac{ln(10)-\mu}{\sigma})}{\Phi(\frac{ln(10)-\mu}{\sigma})}$

Comparison Criteria

1. *Ability to faithfully reproduce underlying model parameters*: Using simulated test data generated according to a deterministic equation with known parameters, is the imputation method able to faithfully replicate those underlying parameters when artificial censoring and missingness is introduced?
2. *Predictive Bias*: $E(\hat{Y} - Y)$, where $\hat{Y}$ is the predicted value and Y is the true value.
3. *Predictive Variance*: $Var(\hat{Y} - Y)$
4. *Convergence Properties*: Does the imputation method converge to any stable state?
5. *Computational Efficiency*: Is the algorithm scalable to larger data sets?

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