Monte Carlo study: t-Student vs Normal regression model with Bayesian inference

Summary

The traditional estimation by the classical model in the presence of outliers can be inefficient, which leads to bad inference and poor decision making. The t-Student distribution is a good alternative in these cases, therefore, the objective of the present work was to evaluate the robustness of the Bayesian regression model with the presence of atypical values using the t-Student distribution comparing to the normal distribution in a Monte Carlo study. It was concluded that the Bayesian linear regression model with t-Student errors is robust in the presence of atypical values that in comparison assuming normality in the errors.

The parameters can be obtained using frequentist inference maximizing the log-likelihood function or can be obtained using Bayesian inference generating a sample of the posterior distribution and the analyzing. In this article the Bayesian inference is performed.

1 Introduction

In real problems, the outliers values is common and they can cause serious problems in statistical analyses, in the literature, there are different ways to treat this observations. The simplest way is to exclude the observation from the data set but it can generate a bad inference, another solution is to use a heavy-tailed distribution like t-student, slash or contaminated normal distributions. In this article it is studied the regression model with the presence of atypical values comparing when the error belong a normal distribution and when they belong the t-student distribution.

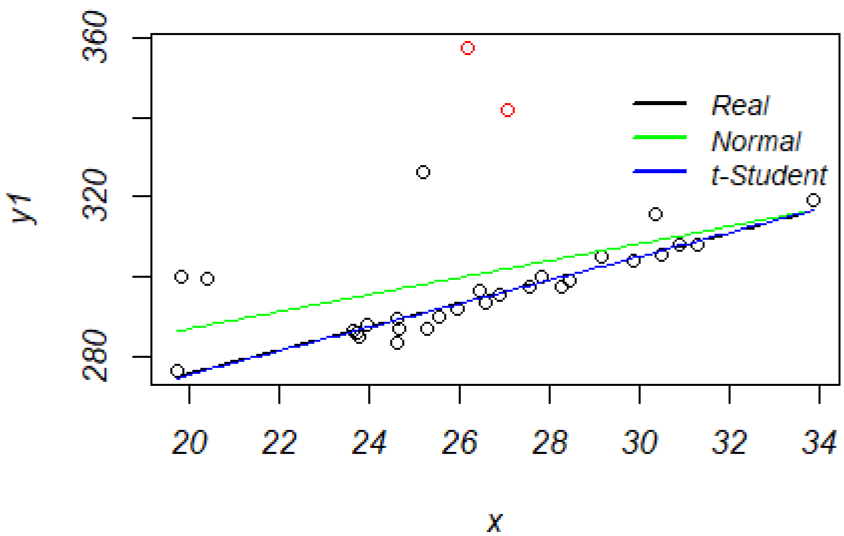

Figure 1: Linear regression model with outliers

The Figure (1) can be observed how the linear regression model with errors t-student has a better fit than the Normal ones.

2 Generating values with atypical observations

The slash distribution is a good option to generate values with atypical observations, in our case it is necessary to evaluate the parameters estimated in the presence of outliers, so the given equation:

$$y_i = x_i^T \beta + \epsilon_i, \qquad \epsilon_i \sim Slash(0, \sigma^2)$$

is a linear regression model with slash distribution on the errors.

3 Classic regression model with Bayesian inference

The equation of linear regression model with normal errors is given:

$$y_i = x_i^T \beta + \epsilon_i, \qquad \epsilon_i \sim Normal(0, \sigma^2)$$

$i = 1, ..., n$, y_i is the target our the variable of interest, $y_i \sim Normal(x_i^T \beta, \sigma^2)$, $x_i = (x_{i1}, ..., x_{ip})^T$ is a vector with explanatory variables for a given observation and $\beta = (\beta_1, ..., \beta_p)^T$ is a vector of the parameters to be estimated.

Considering a sample of size n, the likelihood of the regression model is

$$L(\beta, \sigma^2, y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{1/2 \left(\frac{y_i - x_i^T \beta}{\sigma} \right)^2} \tag{3.1}$$

To estimate the parameters of the regression model, it is defined the prior distribution:

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

The posterior distribution can be obtained as

$$p(\beta, \sigma^2 | Y) \propto \frac{1}{\sigma^{2(0.5n+1)}} \prod_{i=1}^{n} e^{\left(\frac{y_i - x_i^T \beta}{2\sigma^2} \right)^2}$$

To simulate the posterior distribution Gibbs sampling is used, so it is obtained the full conditional distributions:

$$\beta|\sigma^2, Y \sim Normal(U_\beta, \Sigma_\beta)$$

,with $\Sigma_\beta = [\frac{X^TX}{\sigma^2}]^{-1}$, $U_\beta = \Sigma_\beta^{-1}[\frac{X^TY}{\sigma^2}]^{-1}$,

$$\sigma^2|\beta, y \sim GammaInversa(\frac{n}{2}, \frac{\Sigma_{i=1}^n (y_i - x_i^T\beta)^2}{2})$$

.

The Gibbs sampling algorithm consists in

$$\text{Generate } \beta^{k+1} \text{ from } h(\beta|y, {\sigma^2}^k)$$

$$\text{Generate } {\sigma^2}^{k+1} \text{ from } h(\sigma^2|y, \beta^{k+1})$$

for $k = 1, ...10000$.

4 t-Student linear regression model with Bayesian inference

The equation of linear regression model with t-Student errors is given:

$$y_i = x_i^T\beta + \epsilon_i, \qquad \epsilon_i \sim t(0, \sigma^2, v) \tag{4.2}$$

$i = 1, ..., n$, y_i is the target our the variable of interest, $y_i \sim Normal(x_i^T\beta, \sigma^2)$, $x_i = (x_{i1}, ..., x_{ip})^T$ is a vector with explanatory variables for a given observation and $\beta = (\beta_1, ..., \beta_p)^T$ is a vector of the parameters to be estimated.

The representation of (4.2) is:

$$y_i \sim t(x_i^T\beta, \sigma^2, v)$$

Using the propertie of the t-Student distribution, the equation (4.2) can be expressed:

$$y_i|\beta, z \sim N(x_i^T\beta, \frac{\sigma^2}{z}) \qquad z \sim Gamma(\frac{v}{2}, \frac{v}{2})$$

Considering a sample size n, the z_i as latent variables and the degree of freedom, $v(v > 2)$, fixed, so the likelihood with the latent component is given as:

$$L(\beta, \sigma^2, Z, y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{z_i^{1/2}}{\sigma} e^{\,1/2\left(\frac{y_i - x_i^T\beta}{\sigma}\right)^2 z_i} \frac{v/2^{(v/2)}}{\Gamma(v/2)} z_i^{v/2-1} e^{-v/2z_i} \tag{4.3}$$

The prior distribution:

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

The posterior distribution is given as:

$$p(\beta, \sigma^2, Z|Y) \propto \frac{1}{\sigma^{2(0.5n+1)}} \prod_{i=1}^{n} z_i^{1/2} e^{(\frac{y_i - x_i^T\beta}{2\sigma^2})^2 z_i} z_i^{v/2} e^{-vz_i/2}$$

To simulate the posterior distribution Gibbs sampling is used, so it is obtained the full conditional distributions:

$$z_i|y_i, \beta, \sigma^2 \sim Gamma(\frac{v+1}{2}, \frac{1}{2}\left(v + \frac{(y_i - x_i^T\beta)^2}{\sigma^2}\right))$$

$$\beta|Z, \sigma^2, Y \sim Normal(U_\beta, \Sigma_\beta)$$

,with $\Sigma_\beta = [\frac{X^T Z X}{\sigma^2}]^{-1}$, $U_\beta = \Sigma_\beta^{-1}[\frac{X^T Z Y}{\sigma^2}]^{-1}$

$$\sigma^2|Z, \beta, y \sim GammaInversa(\frac{n}{2}, \frac{\Sigma_{i=1}^{n}(y_i - x_i^T\beta)^2}{2} z_i)$$

The gibbs sampling algorithm consists in

$$\text{Generate from } z_i^{k+1} \text{ for } i = 1, ..., n.$$

$$\text{Generate } \beta^{k+1} \text{ from } h(\beta|y, Z^{k+1}, {\sigma^2}^k)$$

$$\text{Generate } {\sigma^2}^{k+1} \text{ from } h(\sigma^2|y, Z^{k+1}, \beta^{k+1})$$

with $k = 1, ...10000$.

5 Results

5.1 Estimated parameters of the classic regression model

Table 1: Posterior mean, SD (standar deviation) and credible intervals (95%) for the parameters in the classic regression model.

Parameter	Posterior mean	Posterior SD	Credible interval 95%
Bo	245.067	191.468	(110.1061; 380.0138)
$B1$	2.057	6.644	(-3.0120; 7.1290)

5.2 Estimated parameters of the t-Student regression model

Table 2: Posterior mean, SD (standar deviation) and credible intervals (95%) for the parameters in the t-Student regression model.

Parameter	Posterior mean	Posterior SD	Credible interval 95%
Bo	216.373	6.0160	(199.4437; 233.3387)
$B1$	2.9596	0.2283	(2.32621; 3.59281)

5.3 Some frequentist properties

Table 3: Frequentist properties of the estimators for the classic and t-Student regression model

Error distribution	Parameter	Bias	EQM	Coverage	Amplitude
Normal	B_0 (216.694)	28.3727	37098.43	0.92	269.9077
	B_1(2.947)	-0.8902	44.4944	0.94	10.141
t-Student	B_0(216.694)	-0.3208	35.9336	0.98	33.8950
	B_1(2.947)	0.0126	0.0517	0.98	1.2666

The Monte Carlo study was done to assess the properties of the parameters of both models. We can see that the linear regression model using the t-student distribution has better properties than the Normal ones.