2024 Summer UTEA Research Report

Project title: “Aspects of Robust Regression”

Supervised by: Nancy Reid & Archer Gong Zhang

Research students: Amanda Ng & Shangkai Zhu

1) Introduction

Traditional regression models, such as Ordinary Least Squares (OLS), have been foundational in predictive modeling. However, these models exhibit significant weaknesses, particularly in the presence of outliers, high-dimensional data, and non-normal error distributions. Outliers can disproportionately influence the estimates, leading to biased and inefficient results. Additionally, in high-dimensional settings, OLS models often struggle with multicollinearity and overfitting, reducing their predictive accuracy and generalizability.

To address these limitations, we proposed using Student-T regression models which provide a more flexible approach to handling deviations from normality and mitigating the influence of outliers. This project explores the implementation of Student-T regression models from both frequentist and Bayesian perspectives. The frequentist approach involves optimizing the likelihood function to obtain MLE for the parameters of the Student-T distribution. Alternatively, the Bayesian approach utilizes MAP on posterior function, which incorporates both priors and likelihood function into the parameter estimation process.

This report aims to provide an overview of

- various Student-T regression models
- optimization issues faced and corresponding solutions
- models' corresponding estimation performance measured by MSE
- confidence interval coverage performances using simulated student-t error, normal error, and outliers data

2) Methods

In the early stage, we verify all theoretical results about student-t regression models from paper "Objective Bayesian analysis for the Student-T regression". Additionally, we run basic simulations to test the performance of Bayesian models using exponential/ Jeffrey's/ reference priors on the degree of freedom parameter.

O. Fonseca, T. C., R. Ferreira, M. A., & Migon, H. S. (2014). Objective Bayesian analysis for the Student-t regression model. Biometrika, 101(1), 252–252. https://doi.org/10.1093/biomet/asu001

BERGER, J. O., & BERNARDO, J. M. (1992). Ordered group reference priors with application to the multinomial problem. Biometrika, 79(1), 25–37. https://doi.org/10.1093/biomet/79.1.25

Wang, M., & Yang, M. (2016). Posterior property of Student-t linear regression model using objective priors. Statistics & Probability Letters, 113, 23–29. https://doi.org/10.1016/j.spl.2016.02.003

2.1) Approach 1: Profile likelihood MLE

This method begins by defining the Likelihood of the Student-T distribution:

$$L(\beta,\sigma,\nu;y,x)=\frac{\Gamma(\frac{\nu+1}{2})^n\nu^{n\nu/2}}{\Gamma(\frac{\nu}{2})^n\pi^{n/2}\sigma^n}\prod_{i=1}^{n}\left\{\nu+(\frac{y_i-x_i^\top\beta}{\sigma})^2\right\}^{-(\nu+1)/2}$$

When estimating $\boldsymbol{\nu}$, we take the nuisance parameters as λ=(β,σ).

We then proceed with constrained MLE on the nuisance parameters by optimizing the log-likelihood function over λ:

$$(\hat{\beta}_\nu,\hat{\sigma}_\nu)=\arg\max_{(\beta,\sigma)}\ell(\nu,\beta,\sigma).$$

Following this, the log profile likelihood is computed by plugging in the constrained MLE to the log likelihood equation:

$$\ell_{\mathrm{p}}(\nu)=\ell(\nu,\hat{\beta}_\nu,\hat{\sigma}_\nu).$$

Finally, the MLE for $\boldsymbol{\nu}$ is obtained by further optimizing the log profile likelihood.

2.2) Approach 2: Adjusted profile likelihood MLE

This method also requires optimizing the log likelihood to obtain constrained MLE on the nuisance parameters. Following this, the log adjusted profile likelihood is computed by plugging in the constrained MLE into the following equation:

$$\ell_{\mathrm{adj}}(\nu) = \ell_{\mathrm{p}}(\nu) - \frac{1}{2}\log\left|j_{\lambda\lambda}(\nu, \hat{\lambda}_\nu)\right|.$$

Notice that log adjusted profile likelihood involved an additional determinant term of the "Observed Fisher information nuisance parameters block". Formulas derived below:

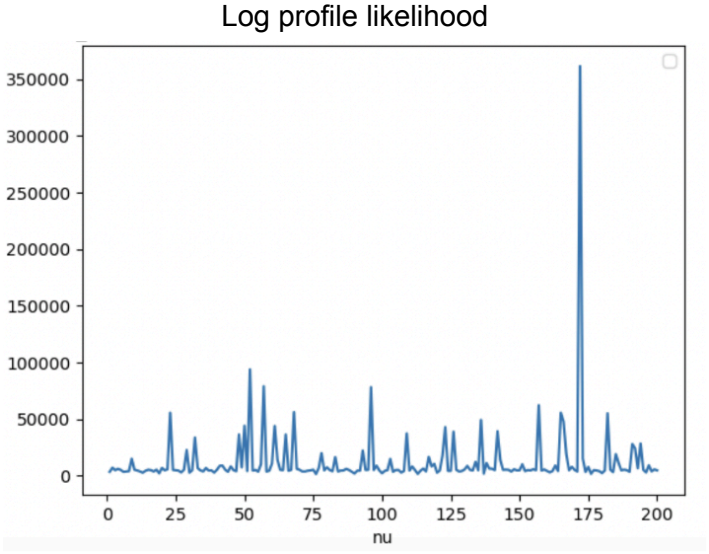
$$j_{\lambda\lambda}(\nu,\lambda) = \begin{pmatrix} -\frac{\partial^2\ell}{\partial^2\beta} & -\frac{\partial^2\ell}{\partial\beta\partial\sigma} \\ -\frac{\partial^2\ell}{\partial\beta\partial\sigma} & -\frac{\partial^2\ell}{\partial^2\sigma} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\nu+1}{\sigma^2}\sum_{i=1}^{n}\frac{x_i x_i^\top}{\nu+z_i^2} - \frac{2z_i^2 x_i x_i^\top}{\left(\nu+z_i^2\right)^2} & \frac{(\nu+1)2\nu}{\sigma^3}\sum_{i=1}^{n}\frac{(y_i - x_i^\top\beta)x_i}{\left(\nu+\left(\frac{y_i-x_i^\top\beta}{\sigma}\right)^2\right)^2} \\ \frac{(\nu+1)2\nu}{\sigma^3}\sum_{i=1}^{n}\frac{(y_i-x_i^\top\beta)x_i^\top}{\left(\nu+z_i^2\right)^2} & -\frac{n}{\sigma^2} + (\nu+1)\sum_{i=1}^{n}\frac{3}{\sigma^4}\frac{(y_i-x_i^\top\beta)^2}{\nu+z_i^2} - \frac{2}{\sigma^6}\frac{(y_i-x_i^\top\beta)^4}{\left(\nu+z_i^2\right)^2} \end{pmatrix}$$

Similarly the adjusted MLE for $\boldsymbol{\nu}$ is obtained by further optimizing the log adjusted profile likelihood.

Challenges and solutions

In both approach 1 and approach 2, we observe optimization issues since the successful convergence rates are low. This is possibly due to the not well-behaved characteristic of log likelihood function (i.e. there are multiple local maximum in the function and the built-in optimization algorithms get stuck at the initial guess point)



Hence, optimizing the likelihood function can be complex due to the presence of multiple parameters and the heavy-tailed nature of the Student-T distribution. To address this, a combination of optimization algorithms is employed: the BFGS algorithm is used in the first step to obtain the constrained MLEs, while the Nelder-Mead algorithm is utilized in the second step to derive the interested parameter's MLE.

Another challenge involves selecting appropriate initial guesses for the optimization process. Poor initial values can lead to suboptimal convergence or failure to converge. To mitigate this, initial guesses are adjusted using least squares estimates for beta and sigma, and the true value for $\boldsymbol{\nu}$.

Lastly, to ensure the parameter $\boldsymbol{\nu}$ and σ remain positive throughout the optimization, we transformed $\boldsymbol{\nu}$ through a logarithmic function, i.e. taking log $\boldsymbol{\nu}$ as the input for all functions, and take the absolute value of the σ, thereby maintaining positivity and enhancing the stability of the optimization process.

2.3) Approach 3: Bayesian model with Jeffrey's prior MAP

In the Bayesian approach, we incorporate prior knowledge into the analysis by specifying priors for all parameters. We use flat priors for betas, an inverse prior for sigma, and the independent Jeffrey's prior on $\boldsymbol{\nu}$, which is defined as

$$\pi^I(\nu) \propto \left\{\frac{\nu}{\nu+3}\right\}^{1/2} \left\{\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \left(\frac{2(\nu+3)}{\nu(\nu+1)^2}\right)\right\}^{1/2}$$

We first compute the profile likelihood with the constrained MLEs which involves optimizing the likelihood function, and multiplying it with the specified priors to obtain posterior distribution. The next step is to optimize the posterior distribution to obtain the MAP estimates.

Challenges and solutions

During the implementation of approach 3, we faced significant runtime issues, particularly in the two-step optimization process and the use of the PyMC MAP function. To overcome these challenges and make the approaches more comparable, we adopted a strategy to optimize all parameters simultaneously within the Bayesian framework. Instead of breaking down the optimization into 2 steps, we formulated a joint density (i.e. multiplying likelihood with priors without integrating it) for the Bayesian model. By optimizing this joint density directly, we achieved a more efficient and streamlined process.

Log-joint density function:

$$\begin{aligned}
&\ell_{MAP}(\beta,\sigma,\nu;y,x)\\
&= n\log\left(\Gamma\left(\frac{\nu+1}{2}\right)\right) + \frac{n\nu}{2}\log(\nu) - n\log\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{n}{2}\log(\pi) - n\log(\sigma)\\
&+ \frac{-(\nu+1)}{2}\sum_{i=1}^{n}\log\left\{\nu + \left(\frac{y_i - x_i^\top\beta}{\sigma}\right)^2\right\} +\\
&0.5\log\left(\frac{\nu}{\nu+3}\right) + 0.5\log\left\{\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right\}
\end{aligned}$$

.

Hessian Matrix:

$$H_{(\beta,\sigma,\nu)} = \begin{pmatrix} \frac{\partial^2\ell_{MAP}}{\partial\beta^2} & \frac{\partial^2\ell_{MAP}}{\partial\beta\partial\sigma} & \frac{\partial^2\ell_{MAP}}{\partial\beta\partial\nu} \\ \frac{\partial^2\ell_{MAP}}{\partial\sigma\partial\beta} & \frac{\partial^2\ell_{MAP}}{\partial\sigma^2} & \frac{\partial^2\ell_{MAP}}{\partial\sigma\partial\nu} \\ \frac{\partial^2\ell_{MAP}}{\partial\nu\partial\beta} & \frac{\partial^2\ell_{MAP}}{\partial\nu\partial\sigma} & \frac{\partial^2\ell_{MAP}}{\partial\nu^2} \end{pmatrix}$$

$$H_{(\beta,\sigma,\nu)} = \begin{pmatrix} \frac{\nu+1}{\sigma^2}\sum_{i=1}^{n} -\frac{x_i x_i^\top}{\nu+z_i^2} + \frac{2z_i^2 x_i x_i^\top}{(\nu+z_i^2)^2} & -\frac{(\nu+1)2\nu}{\sigma^3}\sum_{i=1}^{n}\frac{(y_i - x_i^\top\beta)x_i}{(\nu+z_i^2)^2} & \sum_{i=1}^{n}\frac{z_i(z_i^2-1)\frac{x_i^\top}{\sigma}}{(\nu+z_i^2)^2} \\ -\frac{(\nu+1)2\nu}{\sigma^3}\sum_{i=1}^{n}\frac{(y_i - x_i^\top\beta)x_i^\top}{(\nu+z_i^2)^2} & \frac{n+1}{\sigma^2} + (\nu+1)\sum_{i=1}^{n} -\frac{3}{\sigma^4}\frac{(y_i - x_i^\top\beta)^2}{\nu+z_i^2} + \frac{2}{\sigma^6}\frac{(y_i - x_i^\top\beta)^4}{(\nu+z_i^2)^2} & \sum_{i=1}^{n}\frac{1}{\sigma^3}\frac{(z_i^2-1)(y_i - x_i^\top\beta)^2}{(\nu+z_i^2)^2} \\ \sum_{i=1}^{n}\frac{z_i(z_i^2-1)\frac{x_i^\top}{\sigma}}{(\nu+z_i^2)^2} & \sum_{i=1}^{n}\frac{1}{\sigma^3}\frac{(z_i^2-1)(y_i - x_i^\top\beta)^2}{(\nu+z_i^2)^2} & \bigcirc \end{pmatrix}$$

$$\bigcirc := \frac{\partial^2\ell_{MAP}}{\partial\nu^2} = \frac{n}{4}\left\{\psi'\left(\frac{\nu+1}{2}\right) - \psi'\left(\frac{\nu}{2}\right)\right\} + \frac{1}{2}\sum_{i=1}^{n}\left\{\frac{z_i^2}{1+\frac{z_i^2}{\nu}}\frac{1}{\nu^2} - \frac{z_i^2-1}{(\nu+z_i^2)^2}\right\} - \frac{3}{2}\frac{2\nu+3}{\nu^2(\nu+3)^2} + \frac{1}{2}\frac{\Phi\circledcirc - \circledast^2}{\Phi^2}$$

$$\Phi := \psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}$$

,

$$\circledast := \frac{\partial\Phi}{\partial\nu} = \frac{1}{2}\psi''\left(\frac{\nu}{2}\right) + \frac{1}{2}\psi''\left(\frac{\nu+1}{2}\right) + \frac{2(2\nu^2+9\nu+3)}{\nu^2(\nu+1)^3}$$

$$\circledcirc := \frac{\partial^2\Phi}{\partial\nu^2} = \frac{1}{4}\psi'''\left(\frac{\nu}{2}\right) - \frac{1}{4}\psi'''\left(\frac{\nu+1}{2}\right) - \frac{12(\nu^3+6\nu^2+4\nu+1)}{\nu^3(\nu+1)^4},$$

where $z_i := \frac{y_i - x_i'\beta}{\sigma}$.

Given the Hessian Matrix, we obtain the Observed Fisher Information Matrix,

$$J_{(\widehat{\beta},\widehat{\sigma},\widehat{\nu})} = -H_{(\widehat{\beta},\widehat{\sigma},\widehat{\nu})}$$

2.4) Confidence Intervals

From some previous studies, we believe that the density functions follow χ^2_1 distribution asymptotically. Therefore, it is believed that all parameters should follow Normal distribution, with standard deviation being the square root of the the diagonal elements of J^{-1}, which would be the inverse of the observed fisher information matrix. Then, the confidence interval for beta would be

$$(\widehat{\beta} - 1.96 \times Std, \widehat{\beta} + 1.96 \times Std)$$

where $\hat{\beta}$ is the maximizer (MLE/Adjusted MLE/ MAP).

3) Results

Metric MSE(mean squared error) is calculated as follows: for n repetition and for each $\hat{y}_i$ estimating y_i,

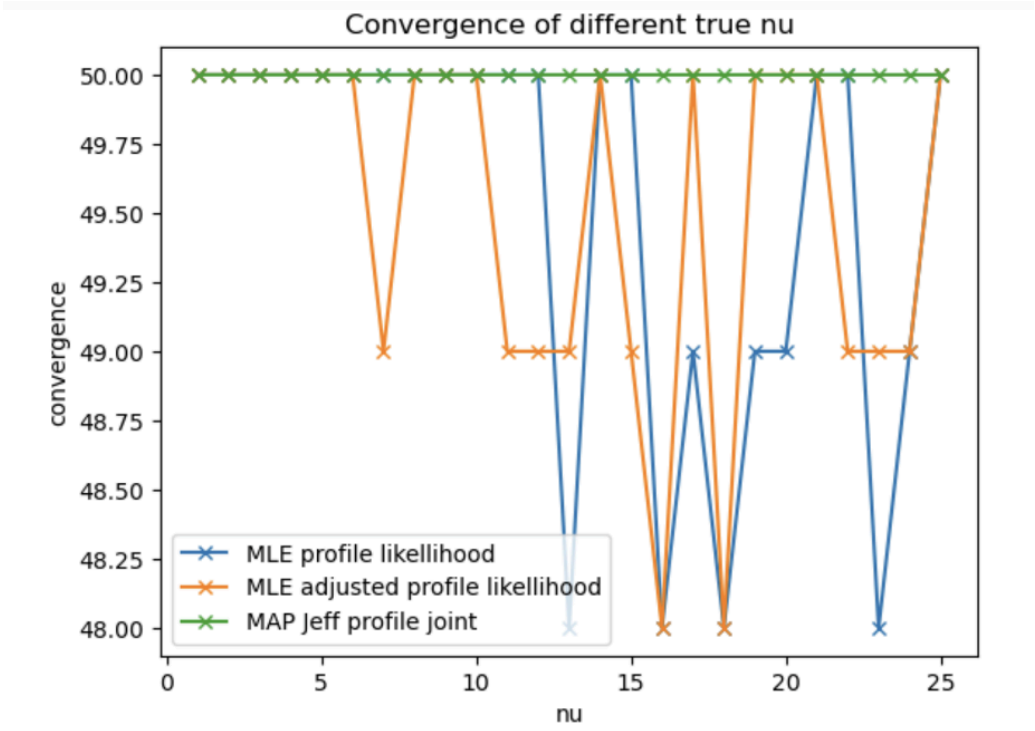
$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)$$

3.1) Convergence rate

Data: Student-t error

$$y = X\beta + \varepsilon$$

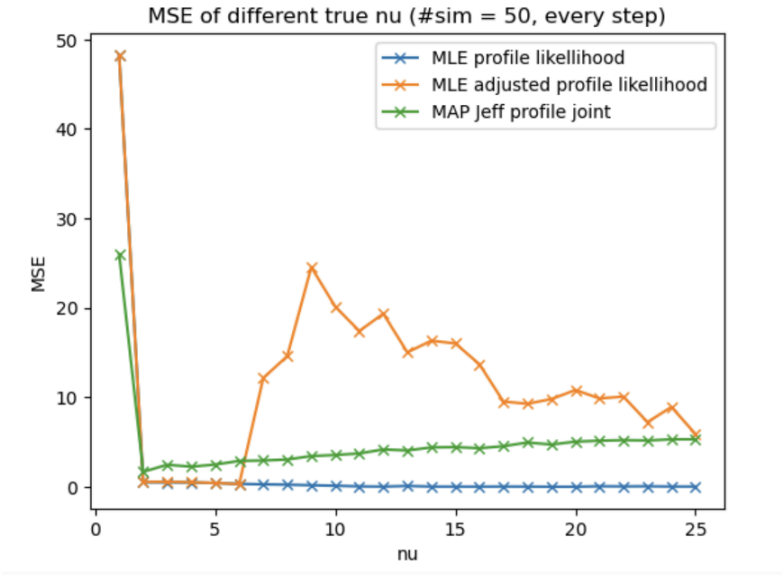
where y $\in R^{n\times 1}$, X $\in R^{n\times p}$, $\beta \in R^{p\times 1}$, and $\varepsilon \in R^{n\times 1}$. The default setting is that β = [2, 1, 0.3, 0.9, 1], n = 200, p = 5, ε follows a Student-t distribution with $\boldsymbol{\nu}$ = 5, σ^2 = 1.5, and $\boldsymbol{\mu}$ = 0.



We conduct 50 repetitions for each $\boldsymbol{\nu}$. As shown in the plot, the Bayesian approach has a promising convergence rate (100%) while the remaining MLE approaches sometimes fail to converge. Nonetheless, all three approaches have relatively high convergence rates.

3.2) Nu MSE

Data: Student-t error (using the same set of data as above: Convergence rate)



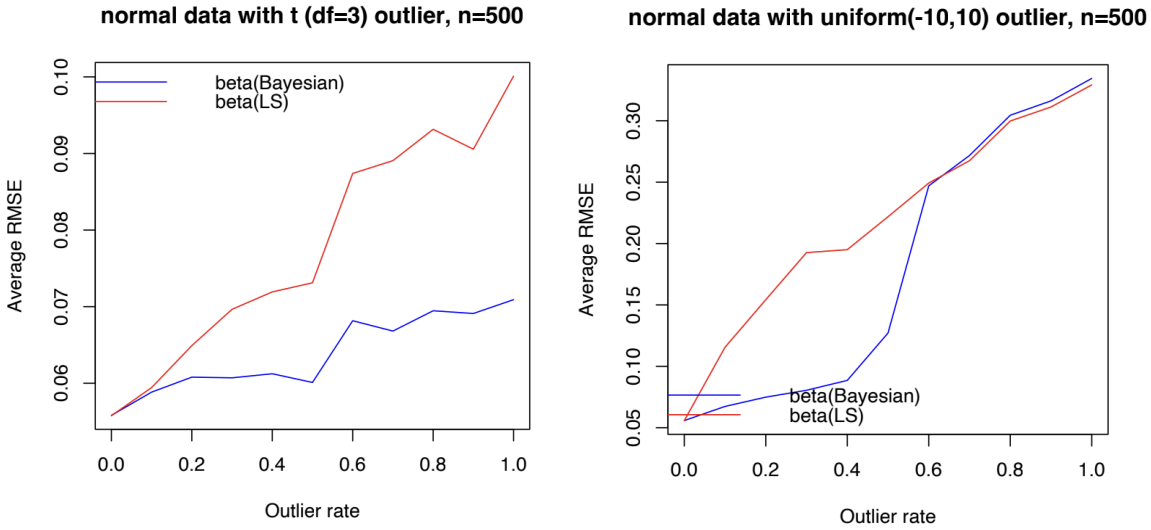
In general, MLE and MAP have better MSE performance than adjusted MLE for estimating ν. Considering the promising convergence rate of MAP, we prefer the Bayesian approach.

3.3) Betas MSE

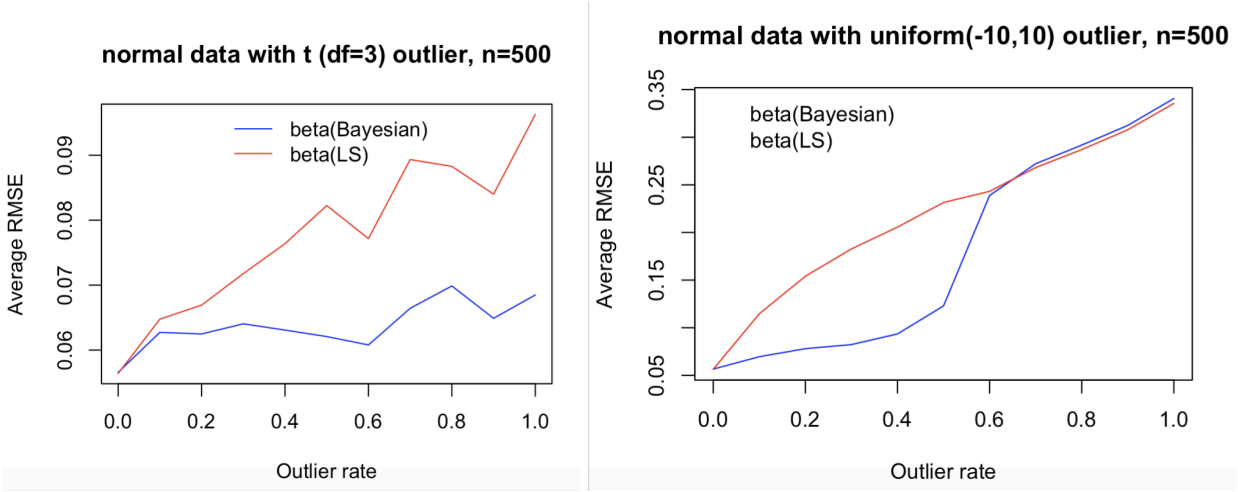
Data: Normal error with outliers which error following other distributions

We compared the averaged MSE of OLS estimates to MAP estimates on βs while varying the outlier rate. When outlier rate = 0, all data comes from Normal(0, 1). When outlier rate = 1, all the data comes from the Student-t(ν = 3, μ = 0, σ^2= 1.5)(left panel) or Uniform(-10, 10) (right panel).

Intercept kept (i.e. 5 betas in total):



Intercept removed (i.e. only left with 4 betas, and we do not add “ones” in X):

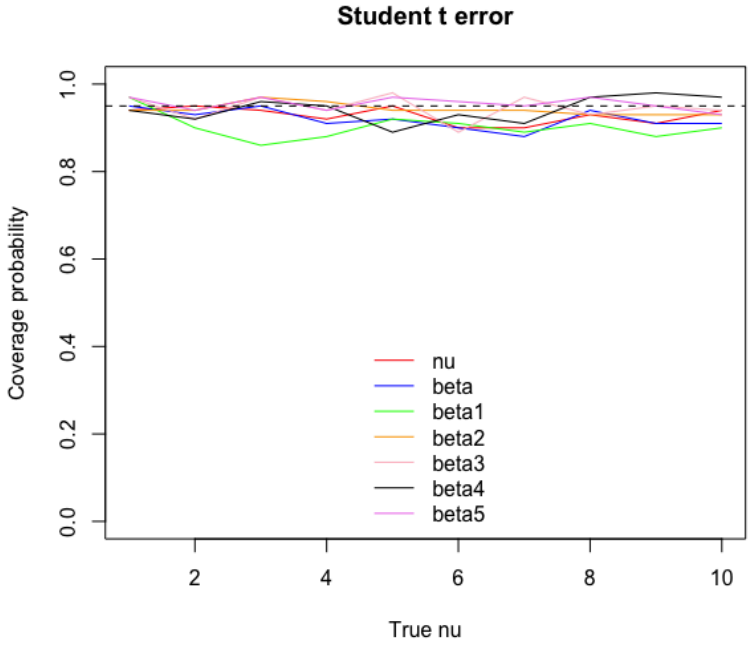


Overall, MAP has better performance in estimating the βs as the outlier rate gets larger.

3.4) Betas Confidence Intervals Coverage Rates

3.41) Construct confidence intervals for each beta separately

Data: Student-t error



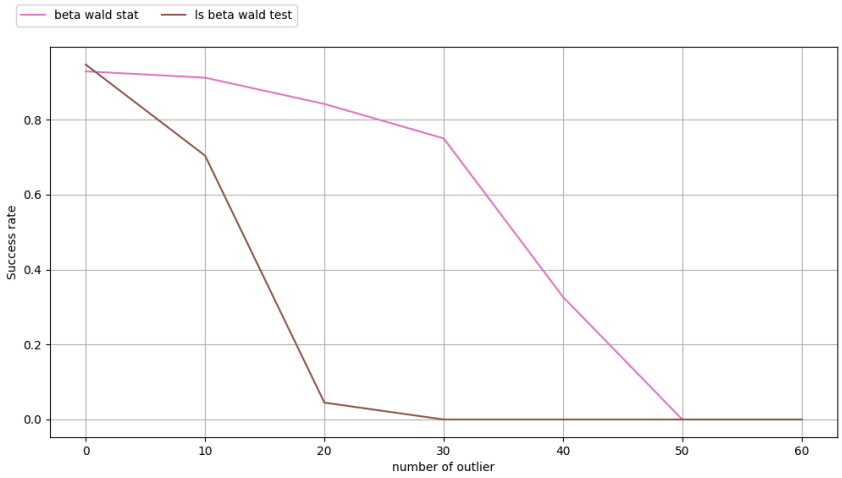
We observe that confidence intervals for all parameters have good coverage rates.

3.42) Construct combined confidence intervals – Wald Statistics

Data: Normal error with Normal outliers (intercept kept)

Error from a mixture of Normal(0, 1.5) and Normal(5, 1):

For a fixed n, α percent of the data would come from Normal(5, 1), and the rest would come from Normal(0, 1.5). α is set to 10 percent of the data for baseline setting.

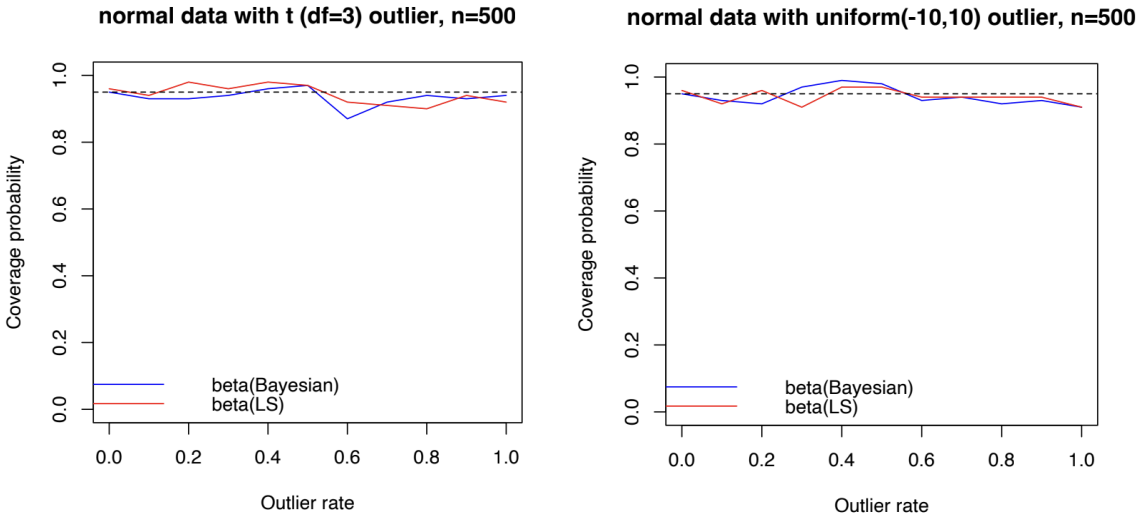


When there are outliers present in the data, MAP confidence interval performs better than OLS confidence interval in capturing the true beta. Similar pattern is also observed in other types of mixture outlier error:

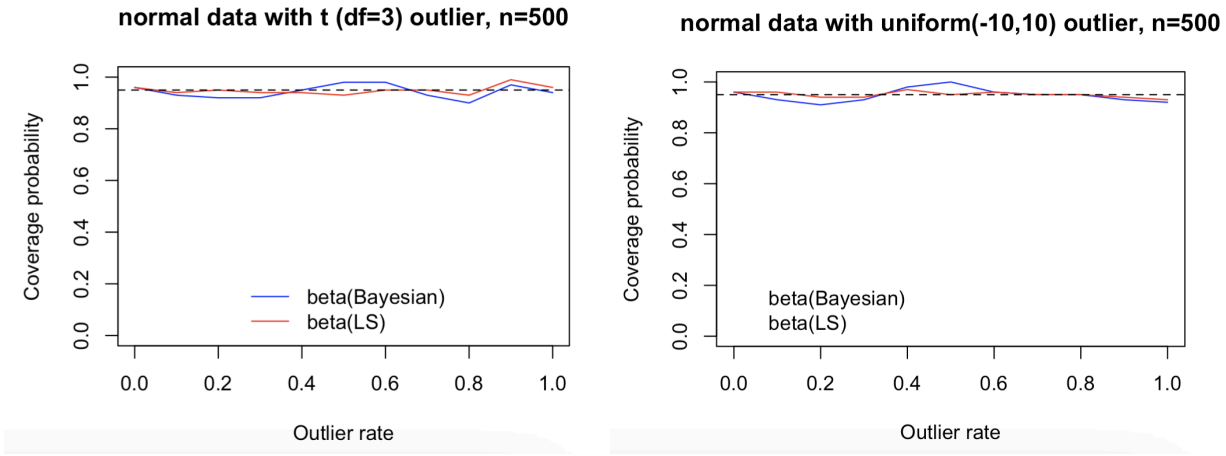
- Student-t(ν = 5, μ = 0, σ^2= 1.5) , Student-t(ν = 5, μ = 5, σ^2 = 1.5)
- Normal(0, 1.5) and Student-t(ν = 5, μ = 5, σ^2 = 1.5)

Additionally, in the mixture of normal errors setting, MAP confidence interval also demonstrates better performance than OLS confidence interval when varying n, p, and σ^2

Data: Normal error with outliers (intercept kept)



Data: Normal error with outliers (intercept removed)



Whether or not intercept is kept, MAP and OLS confidence intervals both exhibit high coverage power regardless of proportion of outliers contained in the data.

4) Conclusion

Accurately estimating the degrees of freedom parameter nu proves to be challenging and has consumed much of our analytical effort. The research findings suggest that Bayesian Student T-regression offers greater robustness compared to OLS, particularly in the presence of outliers or non-normal errors. Some lingering problems to be explored in the future includes:

- Improving convergence rate of models (to reach “always 100%”)
- Negative variance of ν when the number of observations is small or the number of covariates is relatively big
- High dimensional setting