An Applied Analysis of the Bayesian Student-t Regression with Jeffrey's Prior

Jimmy Ting-Yuan Kuo
Department of Statistics, Boston University

Abstract

This paper analyzes a linear regression model where as supposed to using the Gaussian sampling distribution, the student-t distribution is used to robustify the model. The posterior is computed using the Jeffrey's prior and Gibbs Sampler, made possible by a scale-mixture of normals result. An applied analysis on a small and simple dataset set shows that coefficients estimates are indeed more robust towards extreme observations compared to the standard Gaussian model using Jeffrey's prior. I then add artificial noises to the dataset under different settings to compare the student-t and the Gaussian model, confirming the result that posterior of the coefficients are indeed more stable.

Introduction

A problem with assuming a Guassian sampling distribution in a linear regression framework is that the estimates and distribution of the coefficients are not robust towards outliers since the Gaussian distribution is thin-tailed. Moreover, there could be complex variance struture across the individual observations (heteroscedasticity). To overcome the problem, several solutions are provided to robustify the model, particularly in the frequentist literature, such as Huber (1964) and White (1980). To overcome the problem in the Bayesian framework, using the student-t sampling distribution has been proposed. However, since the poserior is difficult or impossible to compute when using the student-t distribution, scales-mixture of normals distribution have been used to treat outliers, first suggested by De Finetti (1961). Linear models with such sampling distribution are subsequently analyzed, notably by Zellner (1976) and West (1984). In this paper, I will first derive the full model utilizing the Jeffrey's prior and the scale-mixture of normals result to obtain the full conditionals, which can be straight-forwardly implemented. I then provide an applied analysis on a US state level dataset on per capital public school expenditure and per capita income in 1979, comparing it with the standard Gaussian error model with Jeffrey's prior. I then add artificial noise to the dataset under various settings, which shows that the distribution of the coefficents are indeed much more invariant under noise.

The Model

The standard linear regression follows the form,

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, 2, 3, ..., n \tag{1}$$

Where $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, y_i is a scalar denoting the "response", $x_i \in \mathbb{R}^{p \times 1}$ where p is the number of predictors, and $\beta \in \mathbb{R}^{p \times 1}$ is the coefficient vector. Since the sampling distribution is assumed to be normal, which is thin-tailed, this model is not robust towards outliers and influencial points, i.e. estimate and inference on β is sensitive towards extreme observations. To "robustify" this model, the sampling distribution can be modified to be fat-tailed. One of the common methods proposed is to take ε_i's as student-t distribution, with v degrees of freem, centered at zero, scaled by σ,

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, 2, 3, ..., n, \quad \varepsilon_i \overset{i.i.d.}{\sim} \text{Student}(v, 0, \sigma^2) \tag{2}$$

In our analysis, we assume v is known. Therefore, under this model $y_i \sim \text{Student}(v, x_i^T\beta, \sigma^2)$, hence the likelihood is given as,

$$L(\beta, \sigma; v, y, x) = \frac{\Gamma((v+2)/2)^n v^{\frac{nv}{2}}}{\Gamma(v/2)^n \pi^{n/2} \sigma^n} \prod_{i=1}^{n} \Big(v + \Big(\frac{y_i - x_i^T\beta}{\sigma}\Big)^2\Big)^{-\frac{v+1}{2}} \propto (\sigma)^{-n} \prod_{i=1}^{n} \Big(v + \Big(\frac{y_i - x_i^T\beta}{\sigma}\Big)^2\Big)^{-\frac{v+1}{2}} \quad (3)$$

Jeffrey's Prior

To estimate this model under the Bayesian framework, a prior over (β, σ^2) needs to be defined. In our analysis we are interested in the Jeffrey prior, $J(\beta, \sigma^2)$. If we were to first relax the assumption that v is know, Fonseca, Ferreira, and Migon (2008) recently propsed Jeffrey's prior for the Bayesian student-t regression for such a model with v unkown.[1] We can specify two types of Jeffrey prior. One, the proper Jeffrey's Rule prior, is given by $J_1(\beta, \sigma^2, v) \propto \sqrt{\det I(\theta)}$, where $I(\theta)$ is the 3×3 Fisher's information matrix over the parameter space $(\theta_1, \theta_2, \theta_3) := (\beta, \sigma^2, v)$, with entry,

$$I(\theta)_{ij} = \mathbb{E}\Big[- \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta \mid y, x)\Big]$$

The second specification is the independence Jeffrey prior, $J_2(\beta, \sigma^2, v) = J_2(\beta)J_2(\sigma^2, v)$, where each marginal prior is computed assuming that other parameters are fixed in the sampling distribution. As shown in Fonseca, Ferreira, and Migon (2008), these two types of priors are different, but they can be expressed in the form,[2]

$$J(\beta, \sigma^2, v) \propto \frac{1}{(\sigma^2)^\alpha} J(v)$$

where $J(v)$ is the "marginal" prior for v. If $\alpha = 1$, then the result reduces to the independt Jeffrey's prior, if $\alpha = (1+p)/2$, where p is the dimension of $I(\theta)$, then the result reduces to the proper Jeffrey's Rule prior.

Since in our analysis we take v as fixed, therefore

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

To see this from the Fisher's informatiotion matrix, reducing the computation in Fonseca, Ferreira, and Migon (2008) to Fisher's Information for standard scale-location parameters (they also parameterize with σ instead), $I(\theta)$ for (β, σ^2) is,

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}\frac{v+1}{v+3}\sum_{i=1}^{n} x_i x_i^T & 0 \\ 0 & \frac{2n}{\sigma^4}\frac{v}{v+3} \end{bmatrix}$$

Therefore,

$$J_1(\theta) \propto \sqrt{I(\theta)_{11} I(\theta)_{22}} \propto \Big(\frac{1}{\sigma^2}\Big)^{(3/2)}, \quad J_2(\theta) = J_2(\beta)J_2(\sigma^2) \propto \sqrt{I(\theta)_{11}}\sqrt{I(\theta)_{22}} \propto 1 \times \sqrt{I(\theta)_{22}} \propto \frac{1}{\sigma^2} \quad (4)$$

$J_2(\theta)$ is by the fact that $J_2(\beta) \propto \sqrt{I(\theta)_{11}} \propto 1$ since $\sqrt{I(\theta)_{11}}$ does not depend on β.

Posterior Distribution

Now that the model is fully specified, the posterior is given by,

[1] Modeling v unknown is a much more difficult problem since only under certain conditions that the posterior is well-defined, see Fonseca, Ferreira, and Migon (2008) for more details

[2] See their appendix section for a detailed derivation

$$
\begin{aligned}
f(\beta, \sigma^2 \mid y, x) &\propto L(\beta, \sigma^2; y, x, v) J(\beta, \sigma^2) \propto \frac{1}{(\sigma^2)^{\alpha}} (\sigma^2)^{-n/2} \prod_{i=1}^{n} \left(v + \left(\frac{y_i - x_i^T \beta}{\sigma} \right)^2 \right)^{-\frac{v+1}{2}} \\
&= (\sigma^2)^{-(n+2\alpha)/2} \prod_{i=1}^{n} \left(v + \left(\frac{y_i - x_i^T \beta}{\sigma} \right)^2 \right)^{-\frac{v+1}{2}}
\end{aligned} \tag{5}
$$

However, there is no closed for solution for the marginals of β and σ^2, nor the full conditionals. To overcome this issue, we can use the well-known result that integrating a normal with variance σ^2 against $\sigma^2 \sim$ inverse gamma will yield a student-t, which is a class of the scale-mixture of normals. Alternatively, this result can be represented in latent variable form,

$$
Z_i \sim \text{Student}(v, 0, \sigma^2) \overset{d}{=} \int_{\mathbb{R}^+} \text{Normal}_{Z_i \mid \lambda_i}(0, \sigma^2 \lambda_i) \times \text{Inv-Gamma}_{\lambda_i}(v/2, v/2) d\lambda_i
$$

Then model (2) can be expressed in the following form with a equivalent marginal sampling distribution, where λ_i is the latent variable,

$$
y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, 2, 3, ..., n \tag{6}
$$

$$
\varepsilon_i \mid \lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \lambda_i), \quad \lambda_i \overset{i.i.d.}{\sim} IG(v/2, v/2) \tag{7}
$$

The interpretation of this Bayesian hierarchical model is that σ^2 determines the overall sampling variability, but λ_i allows for individual heterogeneity, though identically distributed. This is similar to a generalized least squares framework in the frequentist setting. Let $\lambda := (\lambda_1, \lambda_2, ..., \lambda_n)$; therefore, now the "prior" over $(\beta, \sigma^2, \lambda)$ becomes $f(\beta, \sigma^2, \lambda) = J(\beta, \sigma^2) f(\lambda) \propto \frac{1}{(\sigma^2)^{\alpha}} \prod_{i=1}^{n} f(\lambda_i)$. The posterior then is given as,

$$
\begin{aligned}
f(\beta, \sigma^2, \lambda \mid y, x) &\propto L(\beta, \sigma^2, \lambda; y, x) f(\beta, \sigma^2, \lambda) \\
&\propto \frac{1}{(\sigma^2)^{\alpha}} \prod_{i=1}^{n} (\sigma^2 \lambda_i)^{-1/2} e^{-\frac{1}{2\lambda_i} \left(\frac{y_i - x_i^T \beta}{\sigma} \right)^2} \prod_{i=1}^{n} \lambda_i^{-v/2-1} e^{\frac{-v}{2\lambda_i}} \\
&= (\sigma^2)^{-(n+2\alpha)/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{1}{\lambda_i} (y_i - x_i^T \beta)^2} \prod_{i=1}^{n} \lambda_i^{-\frac{v+1}{2}-1} e^{-\frac{v}{2\lambda_i}}
\end{aligned} \tag{8}
$$

Alternatively, to obtain the joint posterior in matrix notation, the linear model can be written as,

$$
Y = X\beta + \Lambda \varepsilon \mathbf{I}
$$

Where $X \in \mathbb{R}^{n \times p}$ is the design matrix, $\Lambda \in \mathbb{R}^{n \times n}$ is a random diagonal matrix with diagonal elements $\lambda_1, \lambda_2, ..., \lambda_n$, $\varepsilon \in \mathbb{R}^{n \times n}$ is also a random diagonal matrix with diagonal elements ε_i, and $Y \in \mathbb{R}^{n \times 1}$ is the response vector. Under this matrix notation, the sampling distribution conditional on λ is $Y \sim \mathbf{MVN}(X\beta, \sigma^2 \Lambda)$, hence the posterior is given as,

$$
f(\beta, \sigma^2, \lambda \mid Y, X) \propto (\sigma^2)^{-(n+2\alpha)/2} e^{-\frac{1}{2\sigma^2} (Y - X\beta)^T \Lambda^{-1} (Y - X\beta)} \prod_{i=1}^{n} \lambda_i^{-\frac{v+1}{2}-1} e^{-\frac{v}{2\lambda_i}} \tag{9}
$$

Full Conditionals

With Jeffrey's prior and scale-mixture of normals representation, the full conditionals can be directly read off. Gibbs Sampler can therefore be easily implemented. Let $\lambda_{-i} := \lambda \setminus \lambda_i$, $RSS_i := (y_i - x_i^T \beta)^2$. Then the full conditionals are,

$$f(\lambda_i \mid \lambda_{-i}, \beta, \sigma^2, y, x) \sim IG\Big(\frac{v+1}{2}, \frac{v+\sigma^{-2}RSS_i}{2}\Big) \tag{10}$$

$$f(\sigma^2 \mid \lambda, \beta Y, X) \sim IG\Big(\frac{n+2(\alpha-1)}{2}, \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - x_i^T\beta)^2}{\lambda_i}\Big) \tag{11}$$

$$f(\beta, \mid \sigma^2, \lambda, Y, X) \sim \mathbf{MVN}\big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\big), \quad \boldsymbol{\mu} = \boldsymbol{\Sigma} \times \big(X^T(\sigma^2\Lambda)^{-1}Y\big), \quad \boldsymbol{\Sigma} = \big(X^T(\sigma^2\Lambda)^{-1}X\big)^{-1} \tag{12}$$

We can see that with Jeffrey's prior, the mean of the full conditional of β is the frequentist Generalized Least Squares estimator for β, and the variance is the variance of the GLS estimator. Utilizing the scale-mixture of normals trick, the student-t model becomes convinient to estimate.

Application

Standard Bayesian Regression with Jeffrey's Prior

To asses how well this "robust" regression performs, it will be compared to the standard Bayesian linear regression ($\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$) with Jeffrey's prior. The proper Jeffrey's rule and the independence Jeffrey's prior are, respectively,

$$J_1(\beta, \sigma^2) \propto \Big(\frac{1}{\sigma^2}\Big)^{3/2}, \; J_2(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Which is identical to student t since it is also a location-scale family with a similar same parametrization given v is known. Under such prior, the posterior marginals can be derived by integrating out the other, after a somewhat tedious calculation[3], let $\hat{\beta} = (X^TX)^{-1}X^TY$ and $s^2 = \frac{1}{n+\alpha-p-1}(Y - X\hat{\beta})^T(Y - X\hat{\beta})$

$$\begin{aligned} f(\beta|X,Y) &\sim \text{Multivariate-}\boldsymbol{Student}\Big(n+\alpha-p-1, \; \hat{\beta}, \; s^2(X^TX)^{-1}\Big) \\ f(\sigma^2|X,Y) &\sim IG\Big(\frac{n+\alpha-p-1}{2}, \frac{n+\alpha-p-1}{2}s^2\Big) \\ &\overset{d}{=} Inv-\chi^2\big(n+\alpha-p-1, s^2\big) \end{aligned} \tag{13}$$

Where $n+\alpha-p-1$ is the degrees of freedom. Hence the marginal of β is centered around the frequentist OLS estimate and scaled by the unbiased estimate of σ^2 if $\alpha = 1$

Data Analysis

To illustrate how a student-t sampling distribution effects the regression, the data we will use is the data set of per capita income and per capita expenditure on public schools by state in 1979, available in R in the `sandwhich` package. Using such a small and simple dataset allows for convinient visualization for illustrative purposes. This data is also analyzed by Fonseca, Ferreira, and Migon (2008), Ferreira and Salazar (2014), Greene (1997), and Cribari-Neto et al. (2000). The dataset has an high-leverage, outlier point from Alaska, apparent from the graph below. The following regression forms will be analyzed,

$$Expenditure_i = \beta_0 + \beta_1 Income + \varepsilon_i \tag{14}$$

$$Expenditure_i = \beta_0 + \beta_1 Income + \beta_2 Income^2 + \varepsilon_i \tag{15}$$

[3] the full marginals for gibbs are $f(\beta|\sigma^2, X, Y) \sim \boldsymbol{MVN}\big((X^TX)^{-1}X^TY, \; \sigma^2(X^TX)^{-1}\big)$ and $f(\sigma^2|\beta, X, Y) \sim IG\Big(\frac{n+2(\alpha-1)}{2}, \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\Big)$

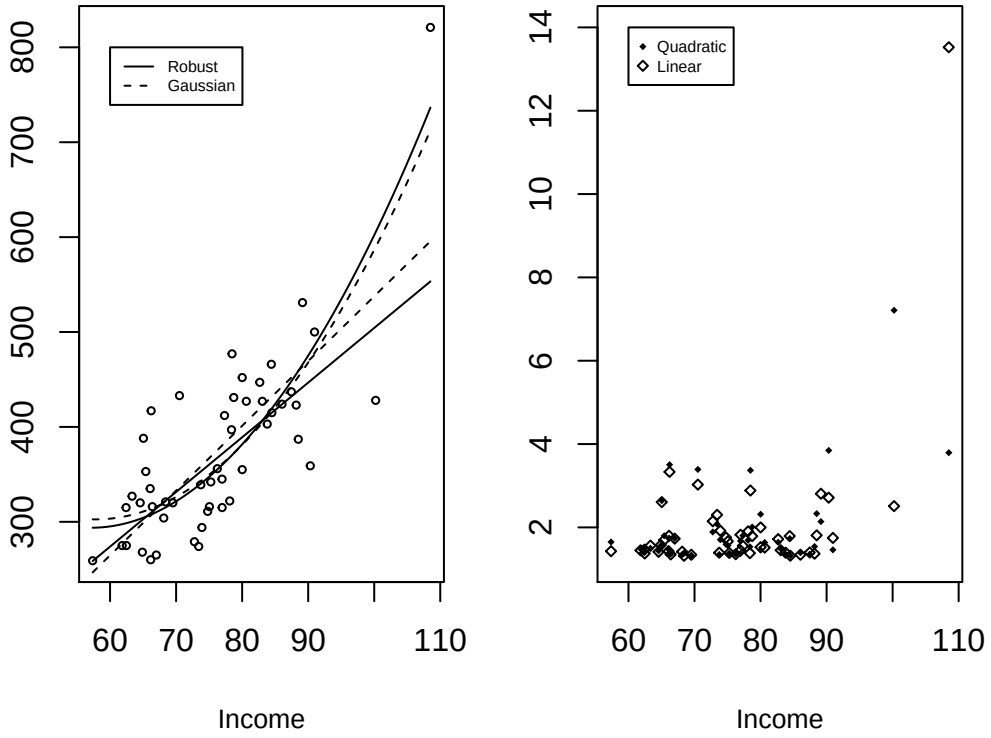

Figure 1: Right: Linear/Quadratic (average) Fits of Gaussian and t errors. Left: Mean Lambda

Table 1: Summary Statistics of Robust Regression

Parameter	**Quadratic**			**Linear**		
	Mean	Median	95% CI	Mean	Median	95% CI
β_0	839.6	903.3	(-271.4, 1579.2)	-71.8	-72.0	(-202.2, 51.5)
β_1	-19.2	-21.1	(-38.2, 11.1)	5.8	5.8	(4.1, 7.5)
β_2	0.17	0.18	(-0.036, 0.29)			
σ^2	2052.9	1974.5	(1176.9, 3371.6)	2156.2	2082.2	(1263.4, 3483.2)

Where ε_i is either normal or student-t.For hyperparameter choices: degrees of freedom $v = 4$ is the conventional choice for the student-t regression in the frequentist setting so it will be used, and the indpendence Jeffrey's prior $\alpha = 1$ will be used. For Gibbs Sampler on the student-t regression, I will run until the effective sample size for each β_i reaches 1000.

The right panel of Figure 1 shows the linear and quadratic fits with the student t and Gaussian erros, where the dotted lines are the Gaussian error fits. We can see that for the linear fit, the Student-t regression is indeed more robust towards the outlier, while for the Quadratic fit, it actually rises more steeply. The left panel of Figure 1 shows the mean λ_i of the independent variable for each observation against income so the effect of λ can be visualized. We see that quadratic has lower λ_i's, as expected since it is more complex. Moreover, all $\lambda_i > 1$, implying that every observation in this dataset displays greater within observation variation than what Gaussian errors would assume.

Table 1 and 2 display the summary statistics. For the Gaussian regression analytically $\mathbb{E}[\beta] = (X^TX)^{-1}X^TY$, $\text{Var}(\beta) = \frac{v}{v-2}s^2(X^TX)^{-1}$, $\mathbb{E}[\sigma^2] = \frac{n+\alpha-p-1}{2}s^2/(\frac{n+\alpha-p-1}{2}-1)$, $\text{Var}(\sigma^2) = \frac{{\theta_2}^2}{(\theta_1-1)^2(\theta_1-1)}$, where θ_1 is the first parameter and θ_2 is the second parameter of the posterior inverse gamma distribution of σ^2, so the summary statistics are based on random draws. We see that for both quadratic fits, the 95% CI of β_i's all cover zero. For the linear fits, the 95% CI of β_1 for the robust regression does not cover 0, while for the gaussian regression it does, which confirms to intuition. The difference is most striking in the variance where the distribution of σ^2 for the robust regression is lower than that of Gaussian errors. The mean of the

Table 2: Summary Statistics of Gaussian Regression

Parameter	**Quadratic**			**Linear**		
	Mean	Median	95% CI	Mean	Median	95% CI
β_0	836.4	837.1	(-3680.1, 5489.0)	-144.7	-147.3	(-1034.0, 754.4)
β_1	-18.5	-18.2	(-135.3, 96.8)	6.8	6.9	(-4.8, 18.2)
β_2	0.16	0.16	(-0.57, 0.89)			
σ^2	3363	3264	(2233.3, 5146.1)	3946	3832	(2621.9, 6006.9)

Table 3: In-Sample Mean Square Error based on the Mean

	Robust	**Gaussian**
Quadratic	3063	3019
Linear	3804	3622

ratio $\frac{\sigma^2_{st}}{\sigma^2_{Gauss}}$ for the quadratic fit is 0.64, with 95% of the distribution nearly below 1, while for the linear fit, $\frac{\sigma^2_{st}}{\sigma^2_{Gauss}} \approx 0.57$ with the 95% CI of $[0.28, 1.04]$. This makes sense as the λ_i's are obsorbing the extra variability.

Table 3 display the in-Sample MSE based on the mean of the posterior of β. Since the Bayes Factor cannot be computed under improper priors,[4] MSE is used for model comparison purposes. We can see that based on this measure, the Robust regression does not in fact perform better than the standard Gaussian error regression.

Mean Zero Gaussian Noise

To see more clearly how robust regression functions, I add noise to the data. I first excluded the high-leverage outlier observation from Alaska, randomly sampled 10% of observations, and then add a $\gamma_i \overset{i.i.d.}{\sim} \mathcal{N}(0, p \times \sigma^2_{Exp})$, where σ^2_{Exp} is the variance of `Expenditure`, to each of the randomly sampled dependent varaible, `Expenditure`, observations. Since now if Alaska is excluded, the linear fit seems more suitable, I therefore fitted the linear fit using both the robust and Guassian regressions on the noisy data. Left panel of Figure 2 shows the regression fits with $p = 0, 0.5, 1, 3$ based on the mean of β, the scatter points are the ones with no noise added, we see that the student-t regression doesn't produce that much of a different result than the standard Gaussian regression. The middle panel shows the mean of λ_i's where a darker color corresponds to a larger p; we see that the λ_i's do not change by much. The right panel shows the Kernel density estimates of β_1 when there is no noise, since β_1 is less sensitive to outliers/influencial points under student-t errors, we see that the distribution is more concentrated towards the middle. The fact that fits from robust regression do not produce much of a different results is indeed not too surprising since the randomly sampled points are not necessarily high-leverage points, i.e. far away from the mean of `Income`, therefore is not guaranteed have a lot of influence on β, moreover these are mean zero noises.

Exponential Noise to Potentially High Leverage Points

To better test how extreme observations effect the two models, I add $\gamma \overset{i.i.d.}{\sim} \exp(\theta)$, $\theta = 0.001,\ 0.0025,\ 0.005,\ 0.01,\ 0.03$ exponential noise to each `Expenditure` with varying survival rate only on observations where `Income`> 89 (89 becaus the average of `Income` ≈ 75) where there are 4 of them. I then model with both the linear and quadratic fits. Figure 3 shows the linear and quadratic fit based on the average of the posterior, the scatter points are uncontaminated data, we can see that the difference is more obvious. From the quadratic fit, we can see that for every θ, the quadratic fit seemed less influence by the extra exponential shock. We can also see how the fit changes as θ decreases (shock increases) to both the Gaussian and Robust model. The linear fit is the most striking, the robust esimates (solid line) more or less remained completely unchanged, even

[4] O'Hagan (1995) propsed the Fractional Bayes Factor for improper priors, but it is hard to compute in the case of the student t regression, MSE is used for model comparison

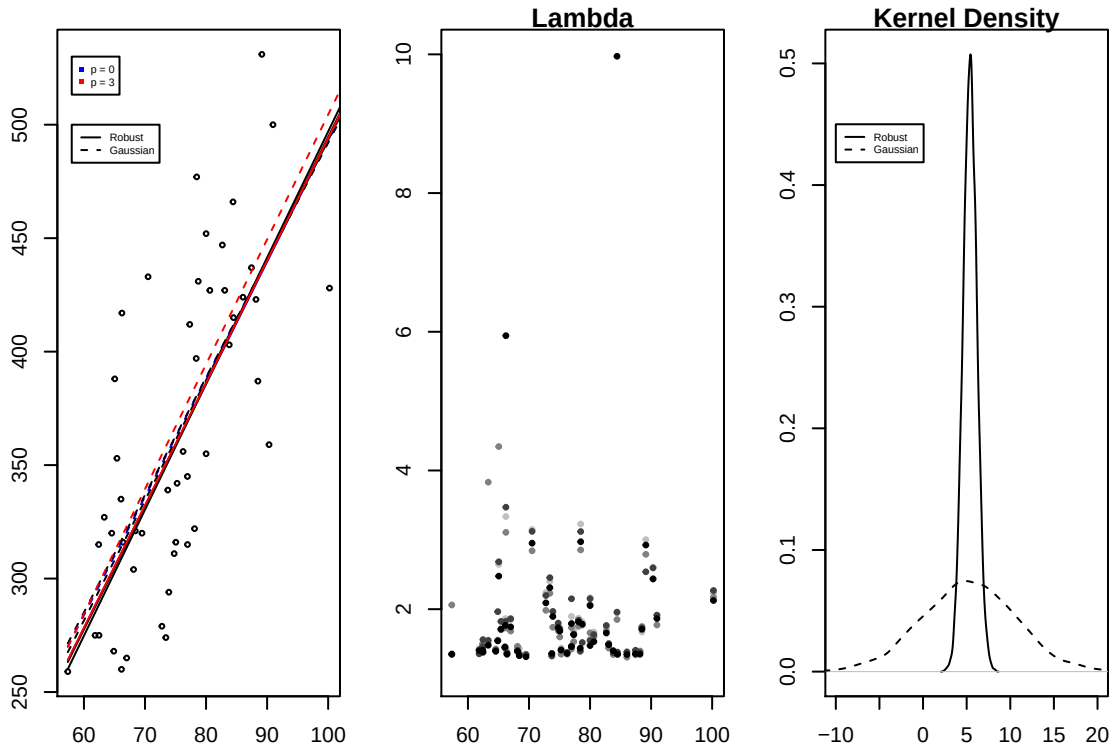

Figure 2: Right: Robust/Gaussian (average) Fits of the Linear Model. Middle: Mean Lambda. Left: Kernel Density of Beta with no noise

when $\theta = 0.001$ hence the expected magnitude of shock is $\frac{1}{\theta} = 1000$. The the mean of `Expenditure` is 364 with a standard deviation of about 70 so the shock is very sizeble. Whereas the Guassian model is heavily effected.

Figure 4 shows the distribution of β_1 of the linear model when $\theta = 0.005, 0.0025, 0.001$. We can see that not only the mean is unaffected as shown by Fig. 3, but the distribution is almost as unaffected. For the Robust regression, the distribution becomes more fat-tailed as θ decreases, but not by much. On the other hand, the posterior of β under Gaussian errors is heavily sensitive to θ, with the overall support much wider than under the robust regression. As θ decreases, the posterior almost becomes flat. As for the estimates on σ^2 (not shown), as identical to the un-modified data, the mean and distribution of σ^2 is much lower in the robust regression than in Gaussian.

Table 4 presents the in-sample MSE under different θ and models; the bottom row is the ratio of Robust MSE and Gaussian MSE. The MSE is calcualted using the mean of the posterior distribution (MSE not computed as a distribution). For the quadratic model, we can see that ratio is about 1, moreover the ratio increases as the shock becomes larger, indicating that the Gaussian model might perform better in the setting of expoential noise to high-leverage points. As for the linear model, the ratio stayed relatively close to 1, and shows a tendency to decrease as the shock becomes larger - opposite to the quadratic case. This shows that the student-t model doesn't neccesarily perform the best when non-normal noise is added to the tail end of the independent variable.

Conclusion

This paper derives the Bayesian student-t regression model with Jeffrey's prior utilizing a scale-mixture of normals result and provides an applied analysis on the potential effects of adding artificial noise to a data set. Comparing to the standard Gaussian error regression also with Jeffrey's prior, the *estimates* of β is indeed much more robust than the Gaussian model across different settings: types of noise added, location of

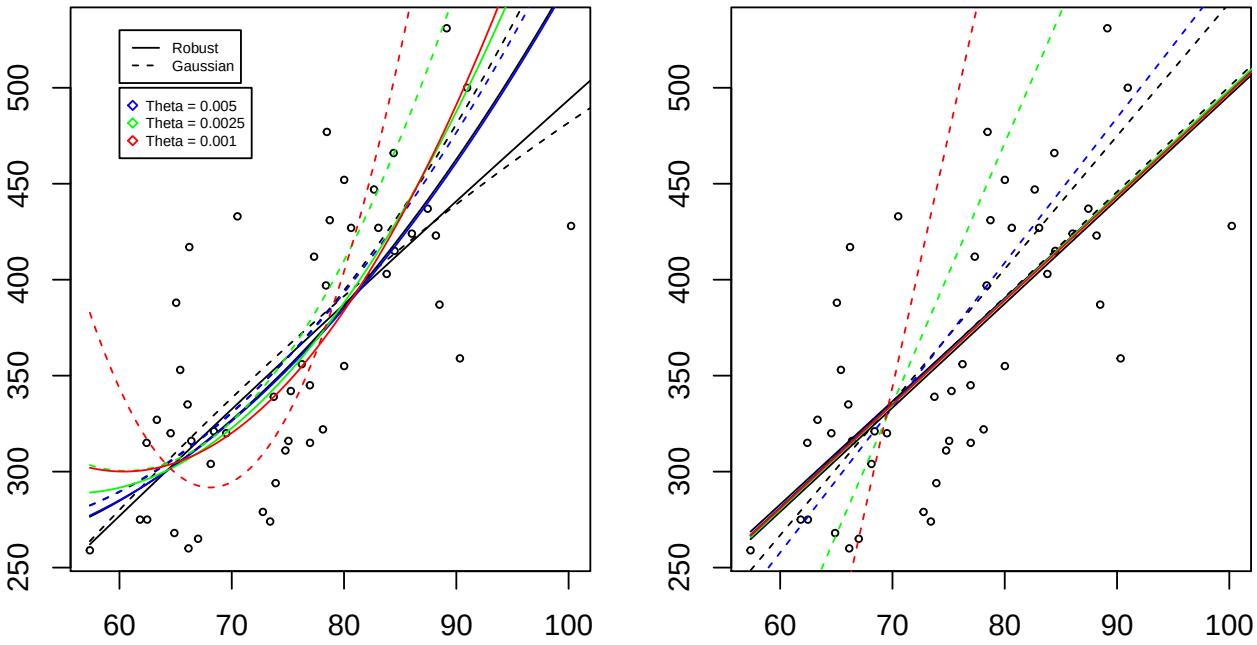

Figure 3: Robust vs. Gaussian Model with Exponential Noise

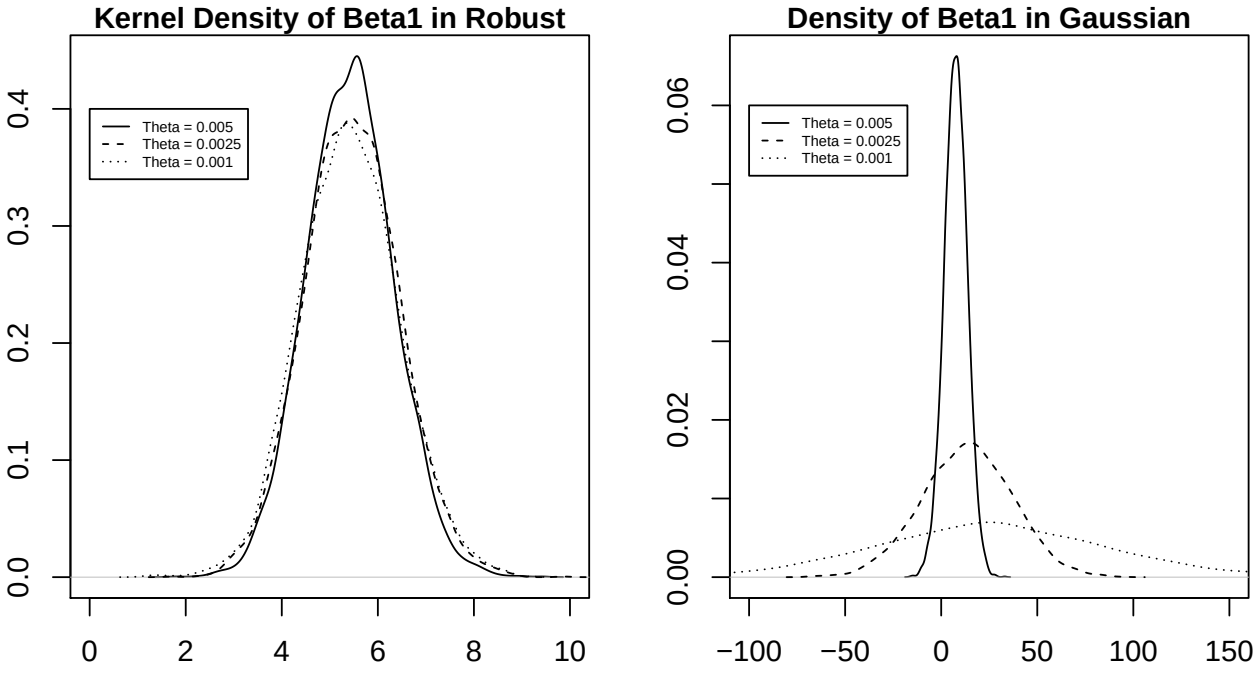

Figure 4: Distribution of Beta1 in the Linear Model

Table 4: In-Sample MSE based on Mean of Beta

Model			**Linear**		
θ	0.03	0.01	0.005	0.0025	0.001
Robust	2529.538	3496.3	4322.569	53387.91	368877.1
Gaussian	2523.249	3630.077	4512.676	62069.99	429570.5
Robust/Gausssian	1.002492	0.9631477	0.9578728	0.8601244	0.8587115

Model			**Quadratic**		
θ	0.03	0.01	0.005	0.0025	0.001
Robust	2474.347	2893.736	3142.189	9432.243	30918.33
Gaussian	2460.319	2819.254	3061.527	8401.551	22285.31
Robust/Gausssian	1.005702	1.026419	1.026347	1.122679	1.387386

noise added, and strenghts of noise added. Estimates of σ^2 is also much smaller in the robust model than in Gaussian, since the extra variability is being absorbed by λ_i's. In the case where fat-tailed noise, like the exponential distribution, is added to high-leverage point observations, the effects of the robust regression is the most pronounced, as expected. The fact that in-sample MSE does not always favor the robust model when non-normal noise is added shows that, though the student-t sampling distirbution provides a more robust result, it does not necessarily provide a better model.

Overall, the effects of the robust regression is likely to be very dataset-dependent. Utilizing the per capital public school expenditure and income dataset, I provide evidence that the robust regression could be more favorable under the linear (non-polynomial) regression setting in the presence of high-leverage points and outliers. However, in my analysis I fix the degrees of freedom, v, on the student-t sampling distribution fix, which does not provide a complete picture of what is going on in the robust regression model. To better understand the strenghts and weekness of the robust regression, besides relaxing the assumption that v is know, several steps can be taken. One would be to generalize the model to high dimensional data and to more diverse datasets. It would also be ideal to perform a predictive analysis and analyze on the out-of-sample prediction errors. This could be an area where the robust regression can perform better, since it sould be more robust towards in-sample randomness.

Works Cited

Cribari-Neto, Francisco, Silvia LP Ferrari, and Gauss M. Cordeiro. "Improved heteroscedasticity-consistent covariance matrix estimators." Biometrika 87, no. 4 (2000): 907-918. \

De Finetti, Bruno. "The Bayesian approach to the rejection of outliers." In Proceedings of the fourth Berkeley Symposium on Probability and Statistics, vol. 1, pp. 199-210. Berkeley: University of California Press, 1961. \

Fonseca, Thaís CO, Marco AR Ferreira, and Helio S. Migon. "Objective Bayesian analysis for the Student-t regression model." Biometrika 95, no. 2 (2008): 325-333. \

Huber, Peter J. "Robust estimation of a location parameter." The Annals of Mathematical Statistics 35, no. 1 (1964): 73-101. \

O'Hagan, Anthony. "Fractional Bayes factors for model comparison." Journal of the Royal Statistical Society. Series B (Methodological) (1995): 99-138. \

West, Mike. "Outlier models and prior distributions in Bayesian linear regression." Journal of the Royal Statistical Society. Series B (Methodological) (1984): 431-439. \

White, Halbert. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." Econometrica: Journal of the Econometric Society (1980): 817-838. \

Zellner, Arnold. "Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms." Journal of the American Statistical Association 71, no. 354 (1976): 400-405.