INTRODUCTION

Dataset

This dataset is the monthly income and expenditure report from the Government of Argentina, Department of Economics. The dataset is downloaded through this website and translated into English.

The dataset included monthly values of 53 economics metrics related to goverment revenue, income and expenditure from 2016 to 2024. This paper aims to reveal important budget allocation insights by exploring the relationship between the total government revenue and the wage expense.

Data cleaning steps

Aggregate total revenue

In the dataset, the government monthly revenue is devided into 10 sub-categories, including VAT net of refunds, earnings, Contributions and contributions to social security, Debits and credits, Personal property, Internal taxes, Fuels, Export duties, Import duties and Other tax revenue. I calculate the predictor variable called "Total revenue" which is the sum of all 10 categories, aggregate by month.

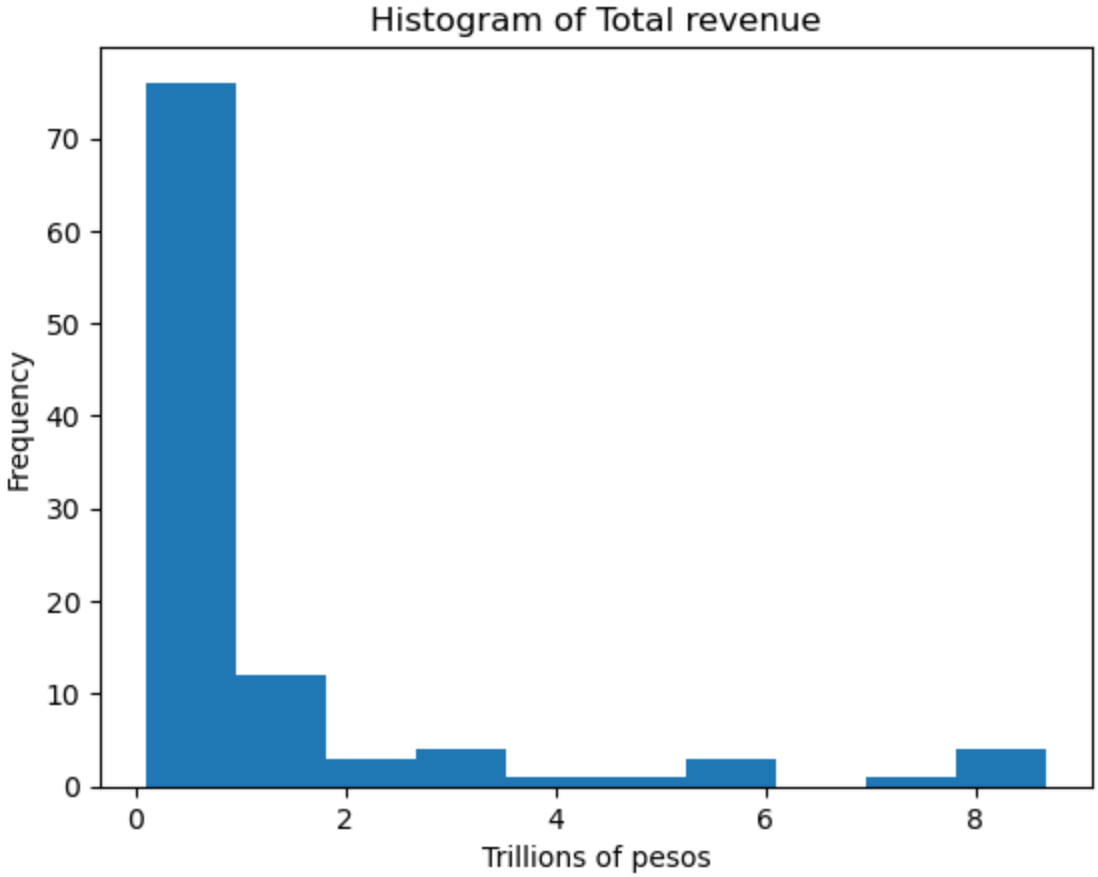
Modify unit of measurement

The unit of measurement in the dataset is millions of pesos. To convert input data into smaller number, I divide each value by 1 million to transform the unit of measurement into trillions of pesos.

Descriptive stats summary

The sample size of the dataset is 105 data points. The predictor variable (Total Revenue) ranges from 0.096 to 8.676 trillions pesos with approximately 70% of datapoints are below 1 trillion pesos. The target variable (Wage Expense) ranges from 0.017 to 1.378 trillion pesos with approximately 75% of datapoints are below 0.1 trillion pesos.

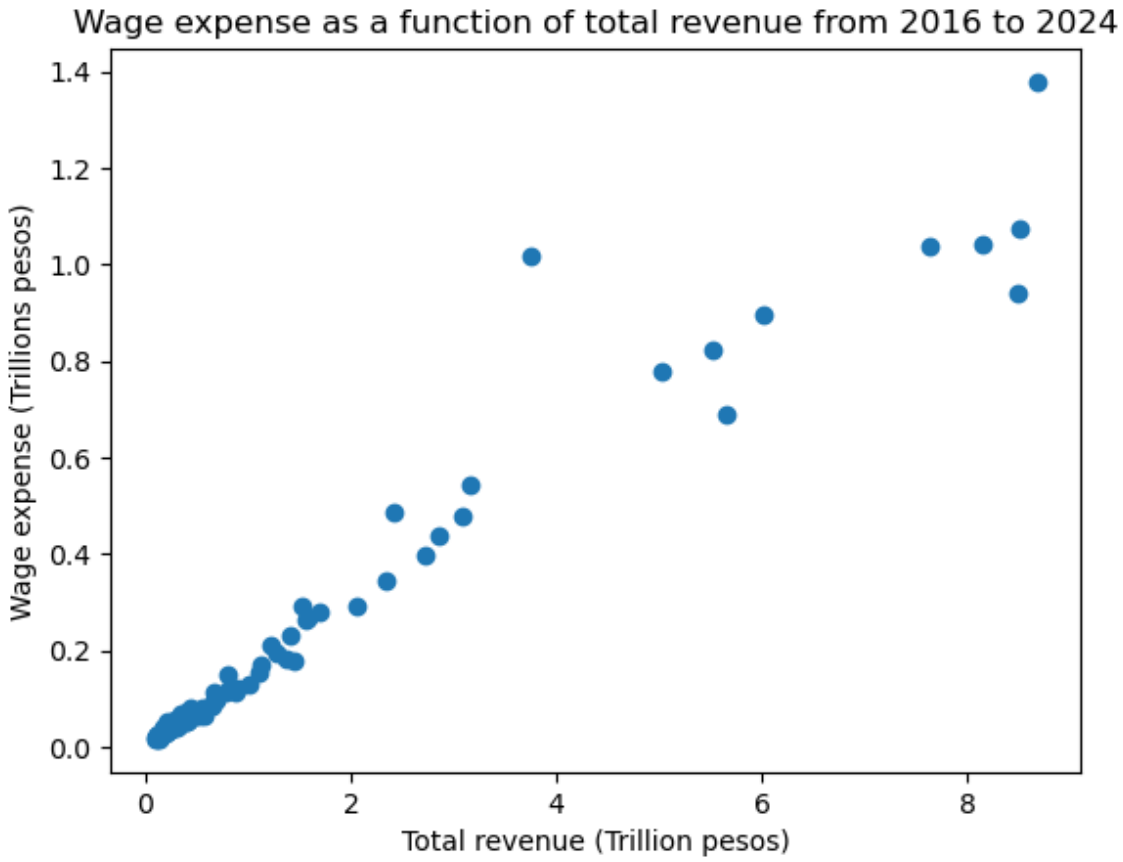
	Count	Mean	Std Dev	Min	25%	Median	75%	Max
Total Revenue (Trillions)	105.0	1.207	1.994	0.096	0.188	0.353	1.129	8.676
Wages (Trillions)	105.0	0.181	0.284	0.017	0.031	0.054	0.171	1.378


Histogram of Total revenue
Frequency
Trillions of pesos
0
10
20
30
40
50
60
70
0
2
4
6
8


Histogram of Wage expense
Frequency
Trillions of pesos
0
10
20
30
40
50
60
70
80
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Scatter plot

Upon looking the scatter plot, we can see a positive linear relationship between total revenue and wage expense. Since both variables have the same unit of measurement, I will not apply standard transformation of the dataset for easier interpretation of the result. I will not apply log transformation to preserve to original unit scale.



CREATE MODELS

I model the relationship using 4 different models: A linear model with Normal likelihood, a quadratic model with Normal likelihood, a linear model with Student T likelihood and a quadratic model with Student T likelihood.

- Linear vs quadratic: This experiments with what polynomial degree with better fit and predict future data
  - Linear Model: Assumes a straight-line relationship between the independent variable x and the dependent variable y which might underfit the data
  - Quadratic Model: Introduces a squared term for the predictor, allowing the model to capture curvature but might overfit if the data doesn't truly follow a curved relationship, especially in cases with sparse data or a lot of noise.

- Normal likelihood vs Student T likelihood: This experiments how sensitive the regression line is to outliers:
  - Normal Likelihood:
    - Assumes residuals are normally distributed around the regression line, with most data points being close to the mean.
    - Sensitivity to Outliers: Since the normal distribution has thin tails, outliers can have a strong influence, potentially pulling the regression line toward them.
    - Adaptation: The parameters of a normal likelihood model are less flexible when handling large deviations, leading to less robust regression results if outliers are present.
  - Student-T Likelihood:
    - The Student-T distribution has heavier tails compared to the normal distribution. This means it is more forgiving of data points that fall far from the mean.
    - Robustness to Outliers: The heavier tails allow the Student-T likelihood to reduce the impact of outliers, maintaining a regression line that better represents the main data trend.
    - Automatic Adaptation: The parameters in a model with Student-T likelihood can adapt to outliers, causing the regression line to remain closer to the non-outlying points, effectively minimizing the outliers' influence on the model.

Normal likelihood model

The linear model relates the independent variable (x - total revenue) with the dependent variable (y - wage expense) by modeling the mean μ_i of y_i as a function of the values in x, through the line that passes through c0 when x_i = 0 and has slope c1

Choosing Priors

C0: Intercept of the linear model that related total revenue to the mean of wage expense

C0 answer the question: What is the expected wage expense, when total revenue = 0. This represents the baseline wage expense for fixed operational human resources which would likely to incur even with minimal revenue.

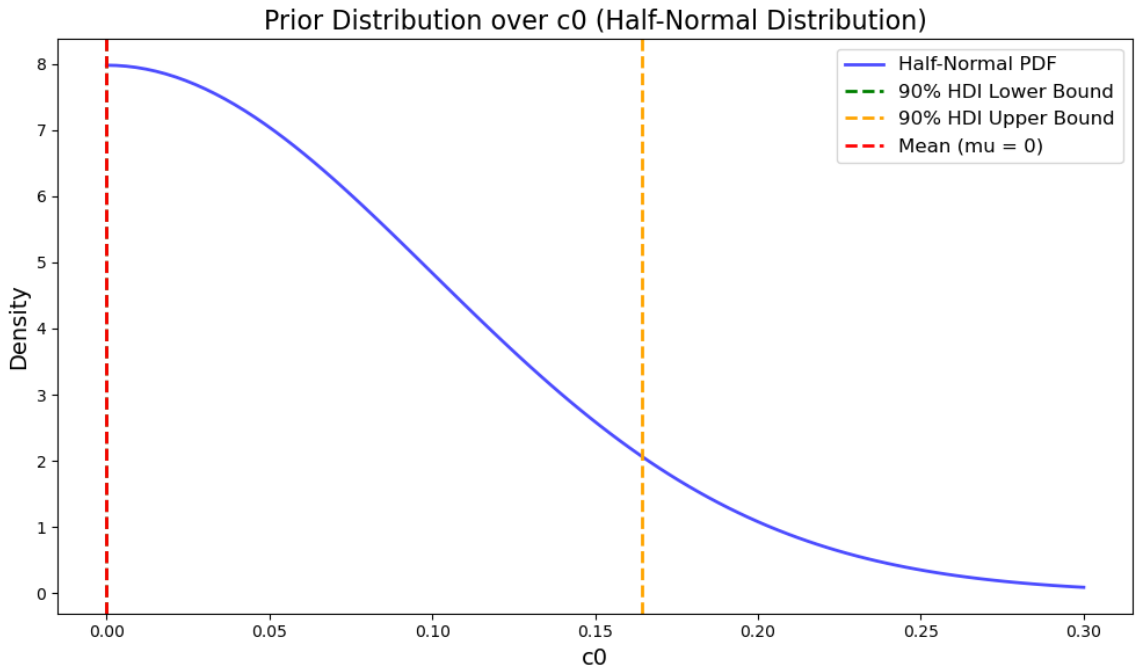
- Wage expense is in trillions pesos, so c0 has to be positive
- When the government has no revenue, it has limited money to spend on wage subsidy (assumption based on economic principles). The distribution of wage expense would center around 0 with limited probability mass on larger values to represent the baseline wage expense.

To estimate the baseline monthly wage expense, I rely on historical data on monthly average salary of workers in public sectors:

- The average monthly salary is around AR$45,200 (Timecamp, 2024)
- The average number of employees in public sector is 3 millions people OECD library, 2023

The estimated baseline monthly wage expense is AR$45,200 × 3,000,000 = 0.135 trillions pesos. Putting 2 requirements together, I choose a Half Normal prior for c0 with sigma = 0.1 and mean = 0.

This is a large enough prior because a 2 standard deviation for the intercept encompase 0.1 x 2 = 0.2 trillions pesos in wage expense. The mode of the Half Normal prior at 0 represent scenarios when the government do not have additional funding sources (loans, one-off funding...) which are variables not capture in "Total revenue" but "Total income" instead. In this case, total revenue = total income = 0, and thus the extreme case: wage expense = 0



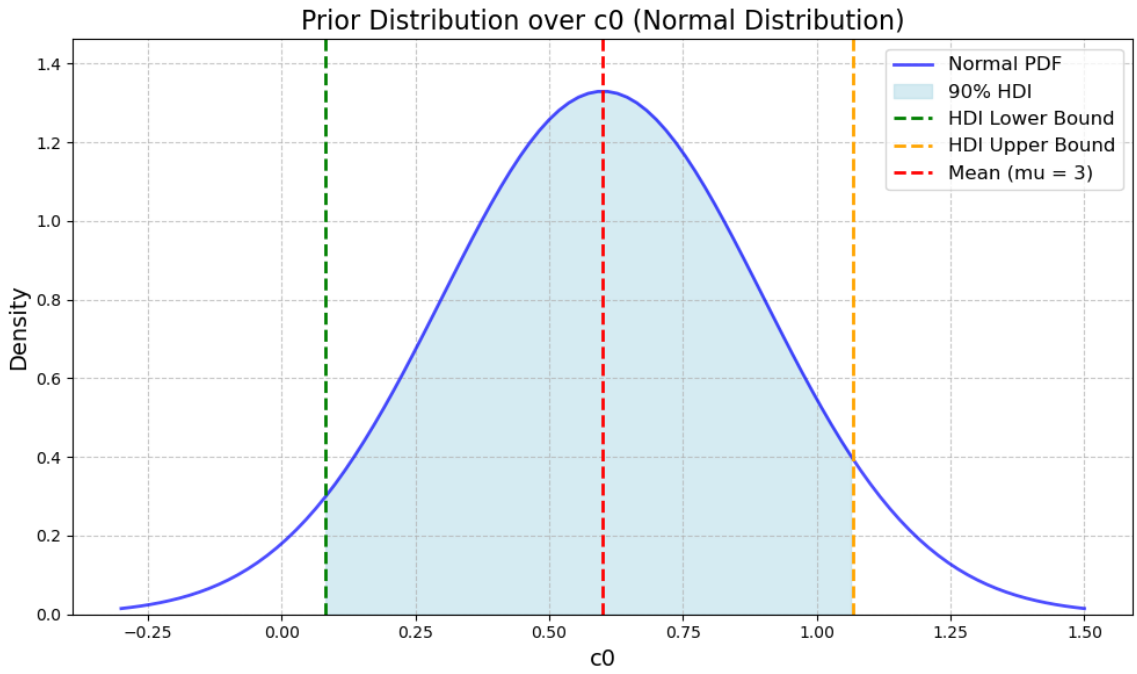
C1: Slope of the linear model that related tax revenue to the mean of wage expense

c1 answers the question: What is the change in expected wage expense, when x_i changes by 1 unit?

Incorporate prior belief: As total revenue increases, it is unlikely that wage expense would decrease (slope < 0). Wage expense would likely to also increases with varying strength. It is unlikely to have a large positive slope because only an approximate 20-30% of total goverment revenue is spent on wage expenditure. I want to allow for some possibility for a slope > 1 since I want to account for situation where the government has to run a budget deficit (spending > revenue), which means that a 1 unit increase in total revenue could correspond to a larger than 1 increase in mean wage expense. This is applicable for Argentina because the government has continuously run a budget deficit from 2009 to cope

with the detrimental effects of various financial crisis and COVID pandemic. In these extreme cases, wage expenses could increase to subsidize businesses/ unemployed citizens.

To reflect this expectation about the slope, I choose a Normal prior with mean = 0.6 and standard deviation of 0.3 to still allow a small probabability mass for the negative values over the left tail, and a larger probability mass for slope > 1 over the right tail (even though still a small proportion). The graph below shows that 90% HDI interval of the prior samples will be in the range of [0.13, 1] which correcty encodes the belief.



Sigma: The uncertainty in the likelihood's mean

The estimate for σ, sigma, informs us of the width of the Normal distribution of wage expense around the mean. For example, for a value of x_i which generate a value of mean μ_i, we have a uncertainty interval for that μ_i where 95% of plausible wage expenses will lie within 2σ away from μ_i.

Since standard deviation is postive, I assign a Uniform prior with bounds from 0 to 0.5. This broad range allows the model to account for the significant variability in wage expense that the total revenue alone cannot explain. Loans and one-off funding, temporal economic cycles, international relations, global crisis, etc contribute to this variability. The prior range for sigma ensures that the model remains flexible enough to capture the inherent unpredictability in the wage expense, while not being so large because wage expense is generally being a fixed proportion of total goverment revenue.

Normal Linear model

Define PyMC model

Likelihood:

$$y_i \sim \text{Normal}(\mu_i, \sigma^2)$$

$$\mu_i = c_0 + c_1 x_i$$

Prior:

$$c_0 \sim \text{HalfNormal}(0.1)$$

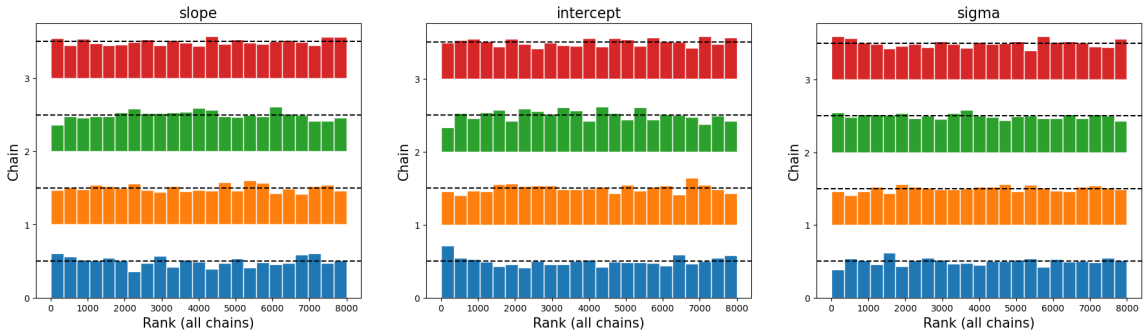
$$c_1 \sim \text{Normal}(0.6, 0.3^2)$$

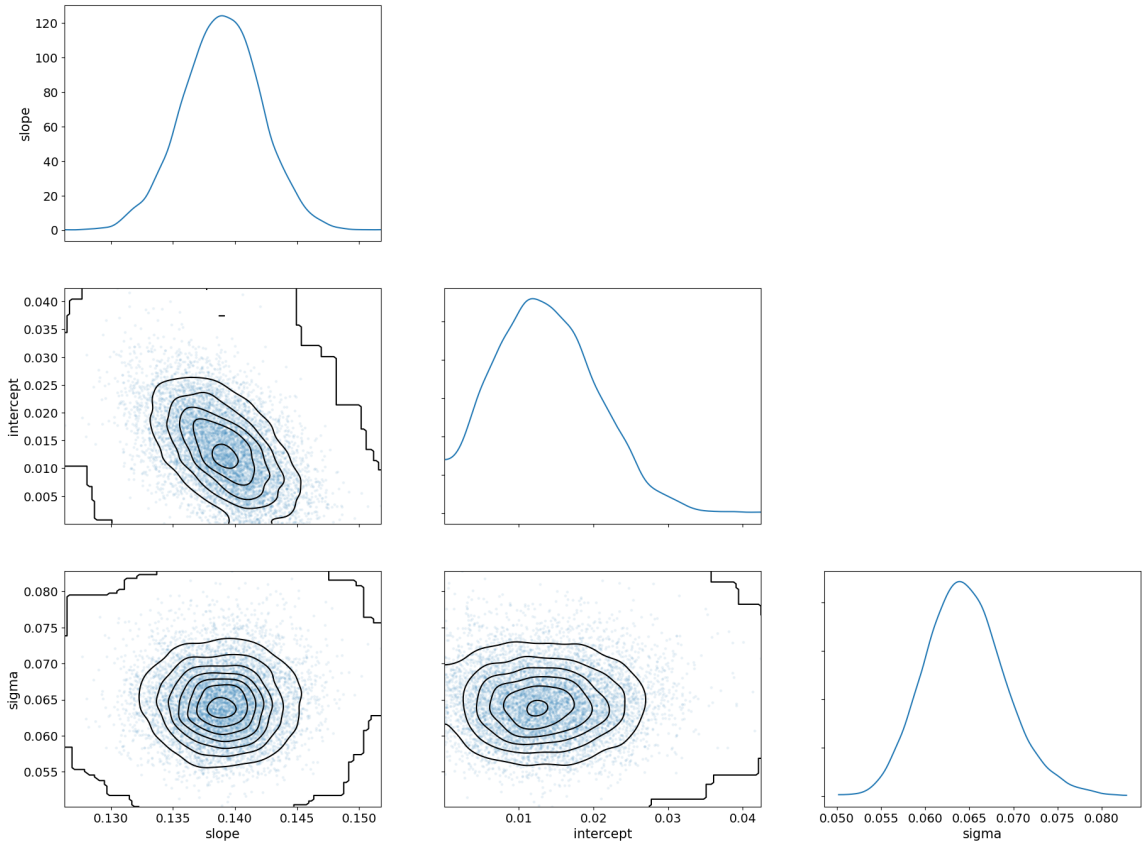
$$\sigma \sim \text{Uniform}(0, 0.5)$$

The chosen priors for c_0 and c_1 is used to deterministically calculate μ_i for each data point x_i through a degree 1 linear function. The normal likelihood is used to draw samples from the gausian distribution of mean μ_i and standard deviation drawn from posterior samples of σ to model how much uncertainty in the μ_i estimate

Sampling diagnostic

For the degree 1 polynomial, the r_hat values are 1, the ess values are high showing both bulk of data and tails are well-sampled from independently, also reflected in the uniformly distribution rank plot histograms which shows convergence. The pair plots also show normal joint distribution of sampled variables.





Normal quadratic model

Define PyMC model

Likelihood:

$$y_i \sim \text{Normal}(\mu_i, \sigma^2)$$

$$\mu_i = c_0 + c_1[0]x_i + c1[1]x_i^2$$

Prior:

$$c_0 \sim \text{HalfNormal}(0.1)$$

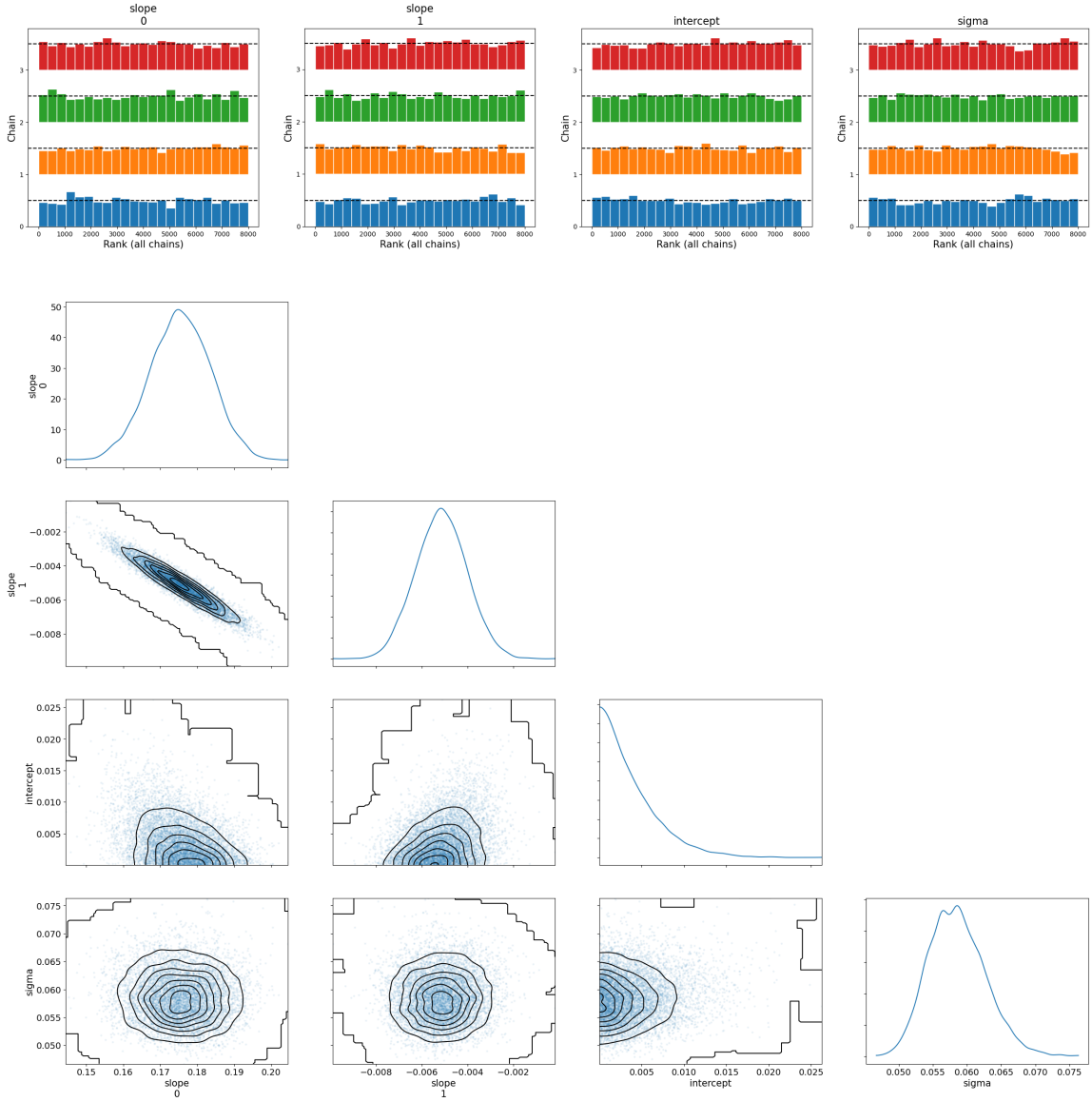
$$c_1 \sim \text{Normal}(0.6, 0.3^2)$$

$$\sigma \sim \text{Uniform}(0, 0.5)$$

In quadratic model, 2 values of c_1 is generated for the slope of x_i and x_i^2 -> the response variabley_i depends on both x_i and x_i^2, allowing for curvature in the relationship.

Sampling diagnostic

The sampler also works well for the same reasons discussed above



Normal cubic model

Define PyMC model

Likelihood:

$$y_i \sim \text{Normal}(\mu_i, \sigma^2)$$

$$\mu_i = c_0 + c_1[0]x_i + c1[1]x_i^2 + c1[2]x_i^3$$

Prior:

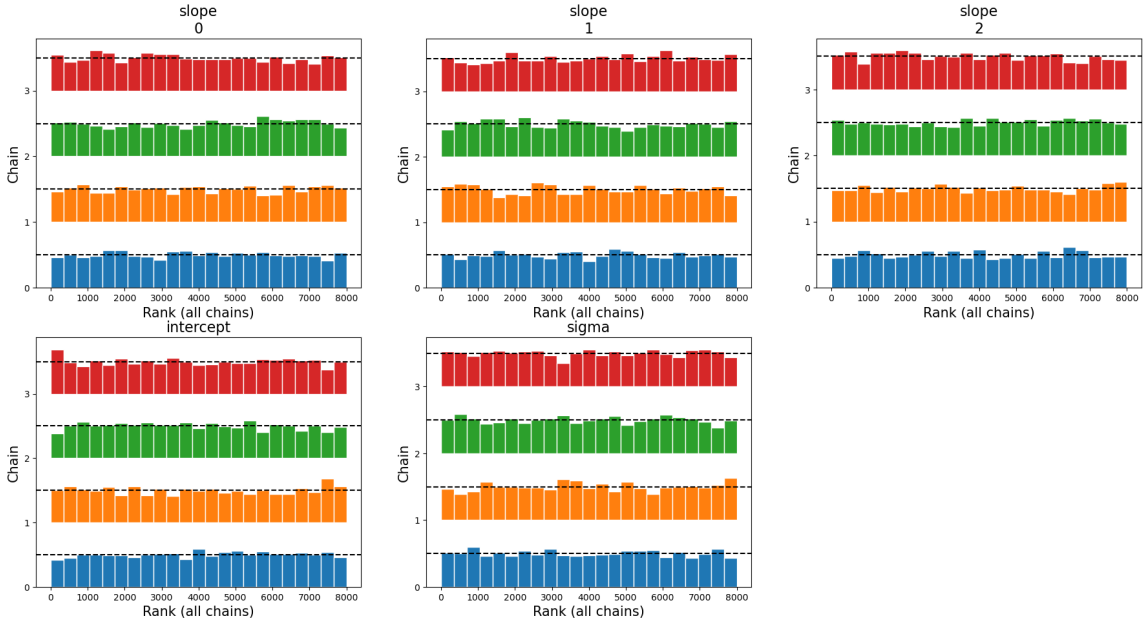
$$c_0 \sim \text{HalfNormal}(0.1)$$

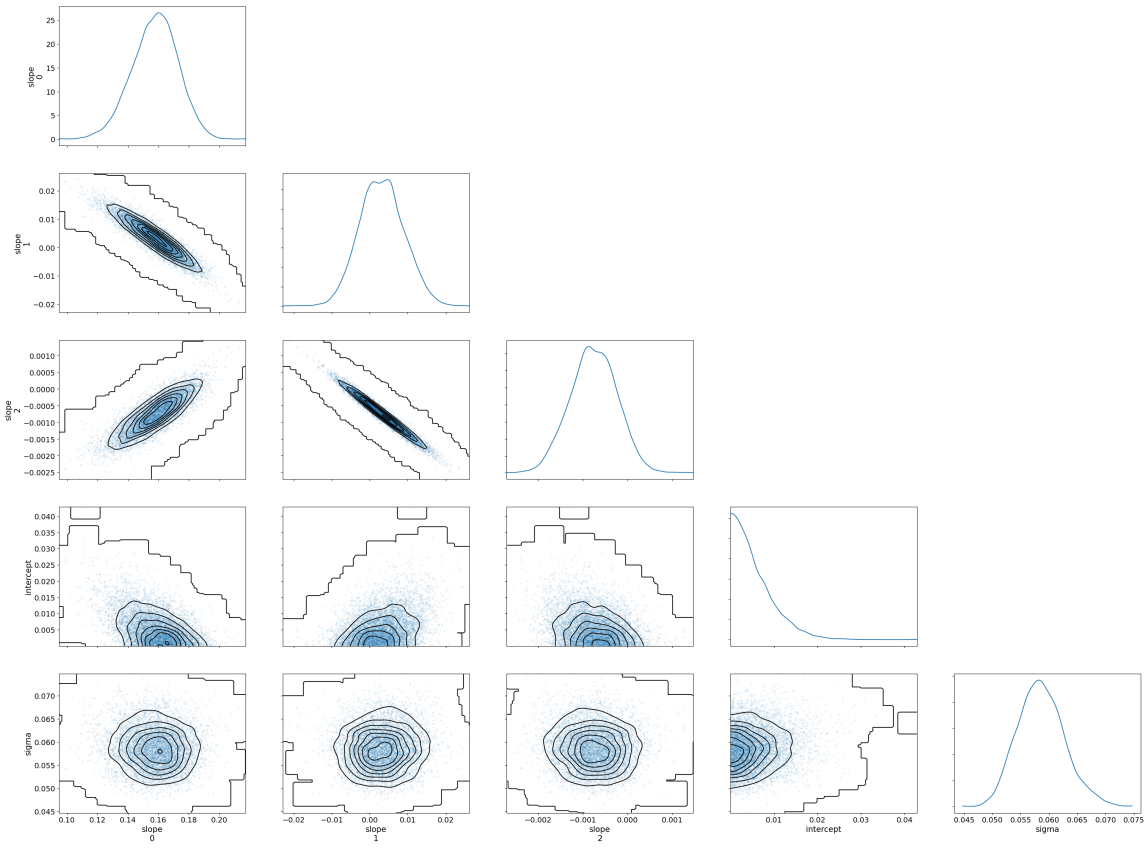
$$c_1 \sim \text{Normal}(0.6, 0.3^2)$$

$$\sigma \sim \text{Uniform}(0, 0.5)$$

3 values of the slope are generated to represent the slopes for the terms x_i, x_i^2, x_i^3 which allows the mean to more complex, asymetric trend.

Sampling diagnostic





Student T likelihood model

The student T likelihood model with heavier tail than normal distribution assumes that the data distribution can produce samples far away from the mean. It has parametr ν that adjusts the lightness of the tails. The larger ν is, the lighter the tails are. When ν ~ 30, the T distribution converges to the Normal distribution but when there are outliers, the model automatically makes ν small, resulting in heavier tails. This allows for a greater likelihood of outliers, so that the σ parameter does not have to increases to fit those outliers anymore. Thus, other parameter estimates also stabilize instead of being pulled away in the direction of the outliers. This flexibility means that the model can accommodate outliers without compromising the overall fit to the majority of the data

Student T linear model

Define PyMC model

Everything is similar to the Normal linear model, except for the parameterization ν. ν is modeled as following a HalfNormal distribution with a standard deviation set to 30, ensures that a positive value support

Likelihood:

$$y_i \sim \text{Student T}(\nu, \mu_i, \sigma^2)$$

$$\mu_i = c_0 + c_1 x_i$$

Prior:

$$c_0 \sim \text{HalfNormal}(0.1)$$

$$c_1 \sim \text{Normal}(0.6, 0.3^2)$$

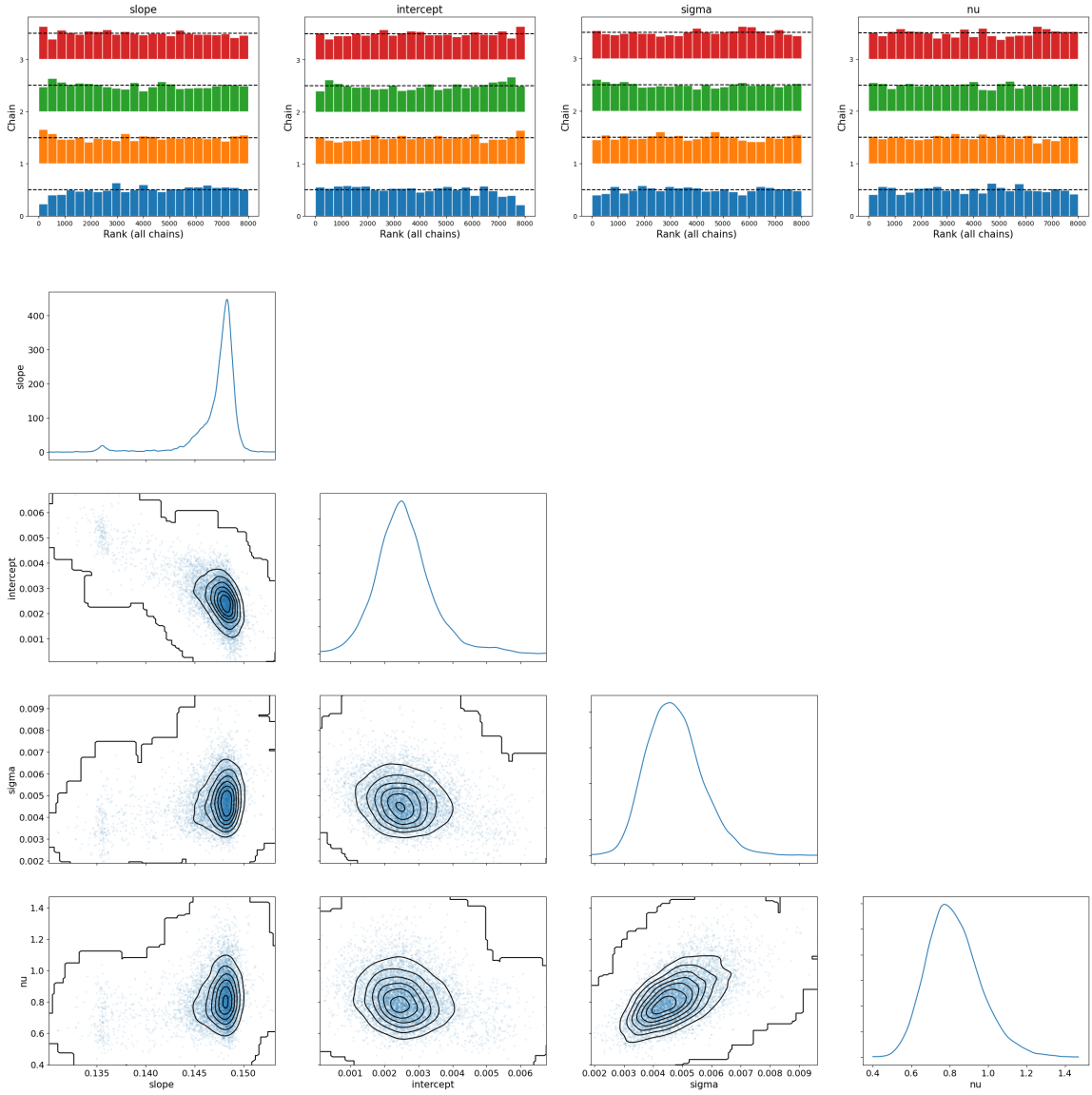
$$\sigma \sim \text{Uniform}(0, 0.5)$$

$$\nu \sim \text{HalfNormal}(30)$$

Sampling diagnostic

The sampler works well for the discussed reason. Notably, we can see in the pairplot the joint distribution between parameter σ and ν shows a positive linear relationship:

- When ν is small, approximately 0.6, σ is also on a smaller range (0.003 - 0.004): This indicates scenario when the Student T distribution adapts to have heavier tail to accommodate outliers without skewing the regression line
- When ν is larger than 1, σ is also on a larger range (0.006 - 0.007): the t-distribution behaves more like a Normal distribution, meaning it assigns less probability to data points far from the mean (lighter tails). In this case, for the model to still fit the data adequately, the scale σ must increase to spread out the distribution and capture outliers.



Student T quadratic model

Define PyMC model

Likelihood:

$$y_i \sim \text{Student T}(\nu, \mu_i, \sigma^2)$$

$$\mu_i = c_0 + c_1[0]x_i + c1[1]x_i^2$$

Prior:

$$c_0 \sim \text{HalfNormal}(0.1)$$

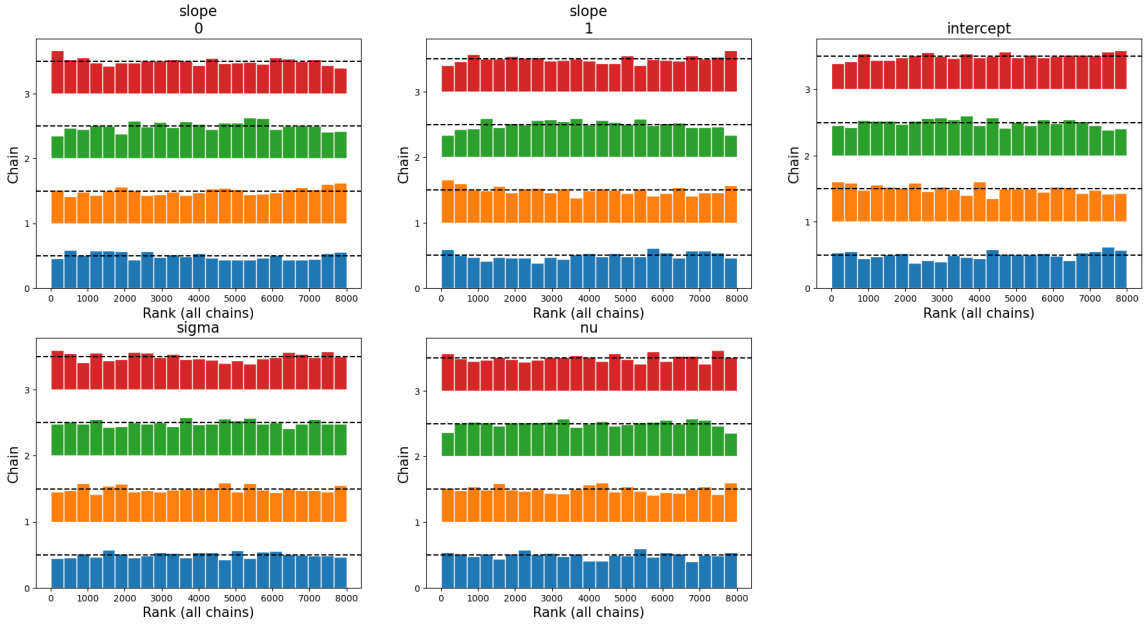
$$c_1 \sim \text{Normal}(0.6, 0.3^2)$$

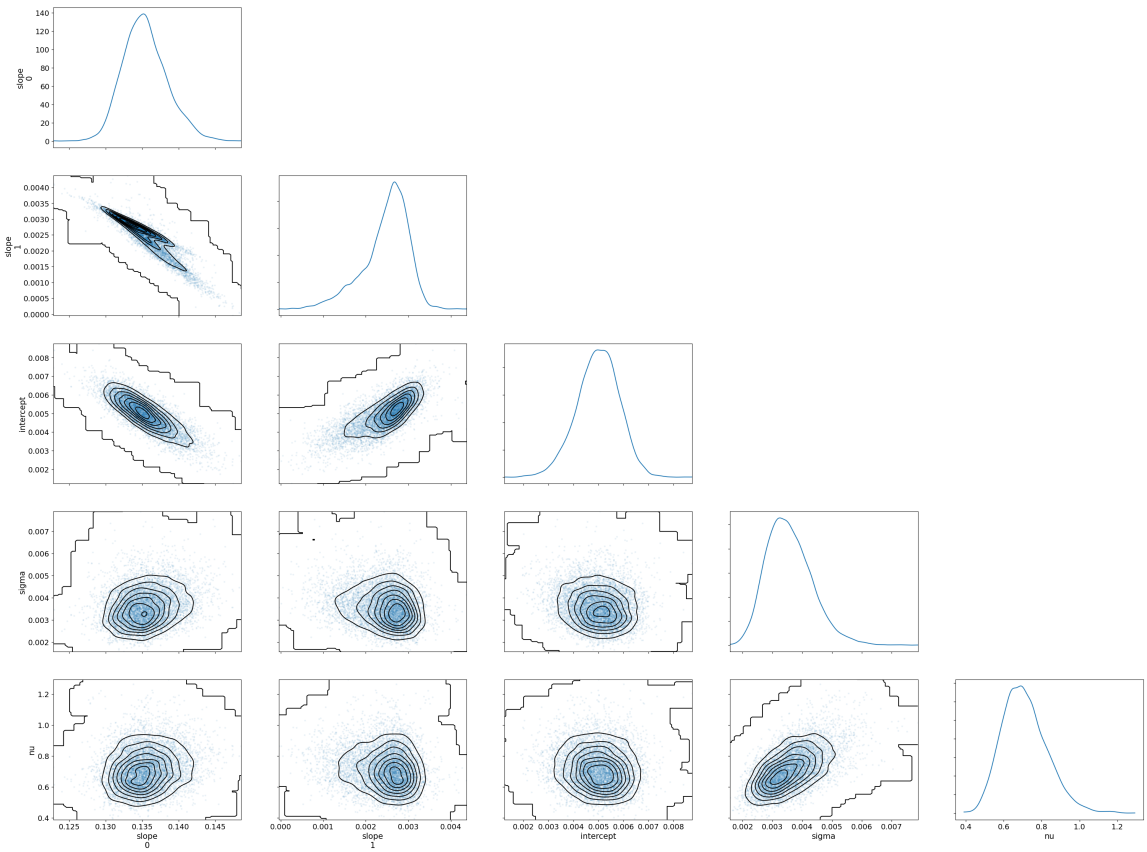
$$\sigma \sim \text{Uniform}(0, 0.5)$$

$$\nu \sim \text{HalfNormal}(30)$$

Sampling diagnostic

The sampler works well.





Student T cubic model

Likelihood:

$$y_i \sim \text{Student T}(\nu, \mu_i, \sigma^2) + c1[2]x_i^3$$

$$\mu_i = c_0 + c_1[0]x_i + c1[1]x_i^2$$

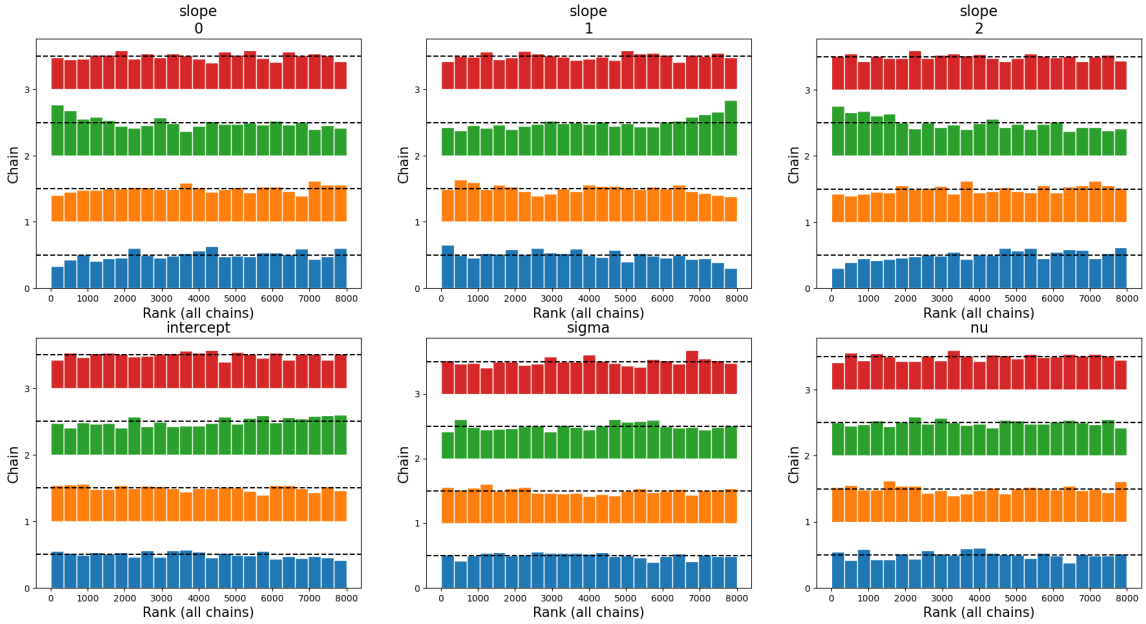
Prior:

$$c_0 \sim \text{HalfNormal}(0.1)$$

$$c_1 \sim \text{Normal}(0.6, 0.3^2)$$

$$\sigma \sim \text{Uniform}(0, 0.5)$$

$$\nu \sim \text{HalfNormal}(30)$$



Posterior distribution of parameters

To understand the robustness and reliability of the model estimates to outliers, I will compare the posterior distributions of parameters (intercept, slope, and sigma) from models using Normal likelihood versus Student T likelihood

Linear models

Marginal posterior distribution of sigma

In the normal likelihood model, the posterior distribution for sigma ranges from [0.05 - 0.08], suggesting a moderate level of uncertainty about the scale of the residuals around the fitted

regression line. However, the range in student T likelihood is significantly lower ([0 - 0.01] range), showcasing no overlap.

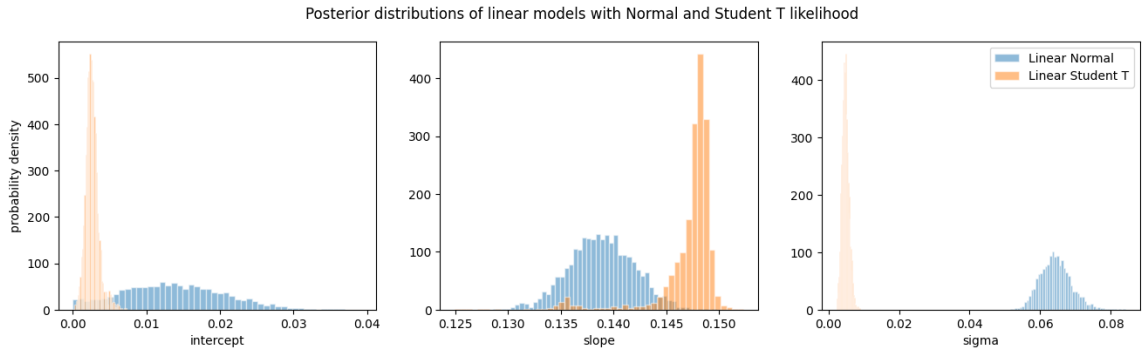
This is inline with the scenarious when outliers are heavily presented, the ν parameter decreases to create heavier tails, and thus σ also decreases. Intuitive explanation for this is that the Student T model allows for the possibility of outliers as part of the data-generating process instead of highly improble data, so it stabilizes the spread of the central tendency, leading to a lower σ estimate.

Marginal posterior distribution of slope

In the Normal Likelihood Model, the slope's posterior distribution ranges from [0.13 - 0.145], while the Student T likelihood shows a left-skewed distribution with very little probability mass between 0.13 and 0.145 but a significant amount in the range of [0.145 - 0.15]. The Student T likelihood, without over-penalizing the clusters of data on the lower right region, interprets the potential slope to be higher than what the Normal likelihood model thinks. In another way, the Normal likelihood model might be shifting the parameter estimation towards the lower end where the outliers are.

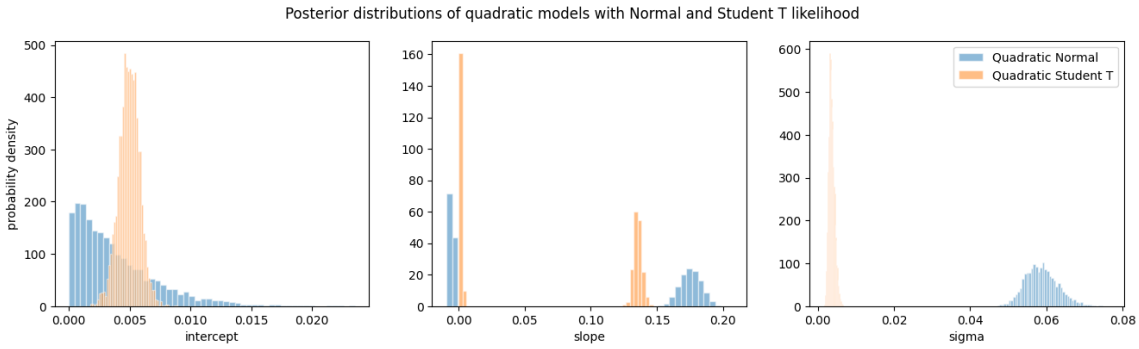
Marginal posterior distribution of intercept

The range for the intercept posteriors also shifts from a wide range [0 - 0.03] to [0 - 0.005]. This is because when the slope posterior increases (as explained above), the intercept must adjust downward to keep the line closer to the data points. This adjustment helps balance the overall fit across the range of x values.



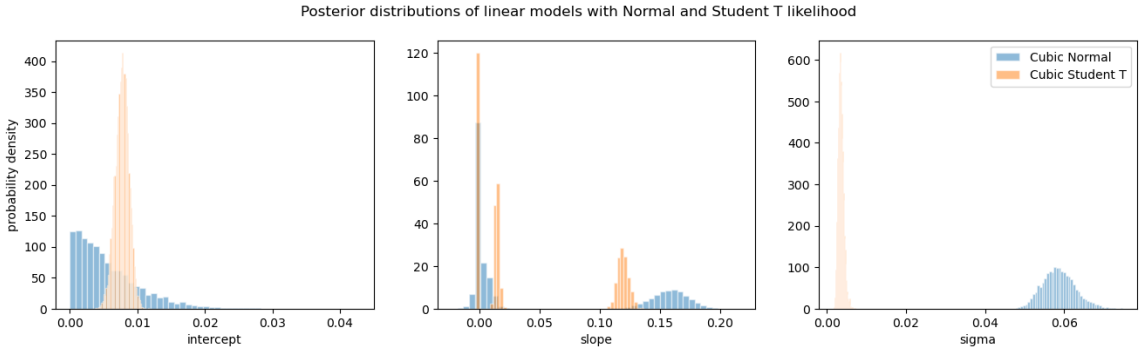
Quadratic models

Being consistent with the above trend, sigma posteriors also shift to the left and slope 0 (linear term) shift to the right. The quadratic term in the Student T likelihood shifts to the left, suggesting a rebalancing of the curve's shape. Under the Student T likelihood, the intercept's posterior is more narrow and bell-shaped, concentrating in a tighter range ([0.002 - 0.008]). This reflects the model's stabilization of the intercept to fit the quadratic regression line.



Cubic models

Similarly to the trend we shown above



Posterior predictive check

The posterior predictive distribution tells us what values we might expect for the dependent variable (wages) given the updated belief about the model parameter uncertainty using the data we have observed (total revenue).

Mathematically, the posterior predictive distribution can be expressed as:

$$(y_{new} \mid x_{new}, data) = \int p(y_{new} \mid x_{new}, \theta) \cdot p(\theta \mid data)\, d\theta$$

Where:

- y_{new} represents new observations of the dependent variable.
- x_{new} are the new values of the independent variable(s).
- data is the observed data used to update our beliefs about the parameters.
- θ represents the model parameters.
- $p(y_{new} \mid x_{new}, \theta)$ is the likelihood of new data given the model parameters.
- $p(\theta \mid data)$ is the posterior distribution of the parameters after observing the data.

This integral combines predictions made by the model across all plausible values of the parameters, weighted by how likely those parameters are given the data.

What is the process behind posterior predictive distribution:

1. **Sampling from the Posterior:**

After observing data, we often obtain posterior distributions for the parameters slope, intercept and sigma (and nu if it is Student T distribution).

Then, For any given value of x_i (total revenue), we first compute the corresponding mean μ_i using the posterior samples for slope and intercept.

The variability in the posterior samples will create undertainty in μ_i estimates, which will give us the **posterior regression line credible interval** (the shaded darker orange region). The averge line among these posterior regression lines will be the MAP - line of best fit, or the **posterior regression mean** (the single darkest orange line)

2. **Sampling Heights:**

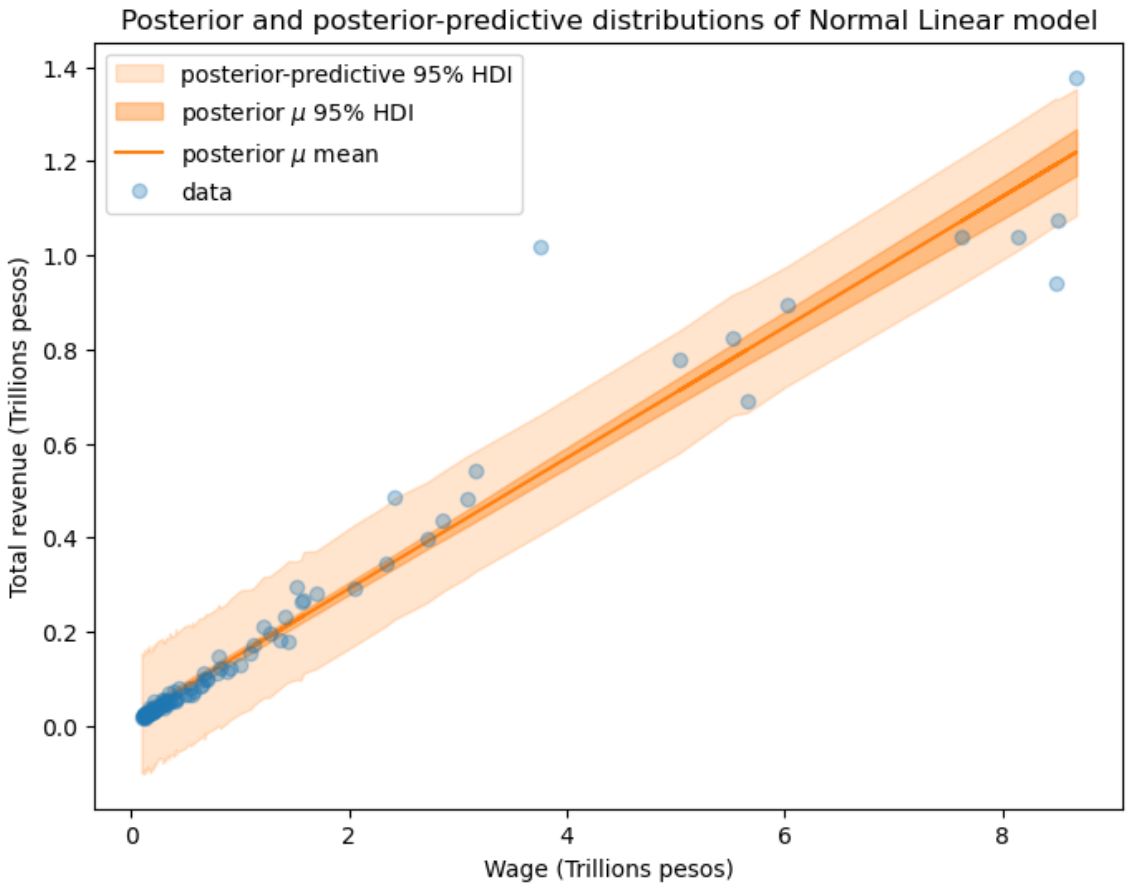
Once we have a mean μ_i for a specific x_i we then sample from a Gaussian distribution centered at that mean with a standard deviation σ sampled from its posterior distribution.

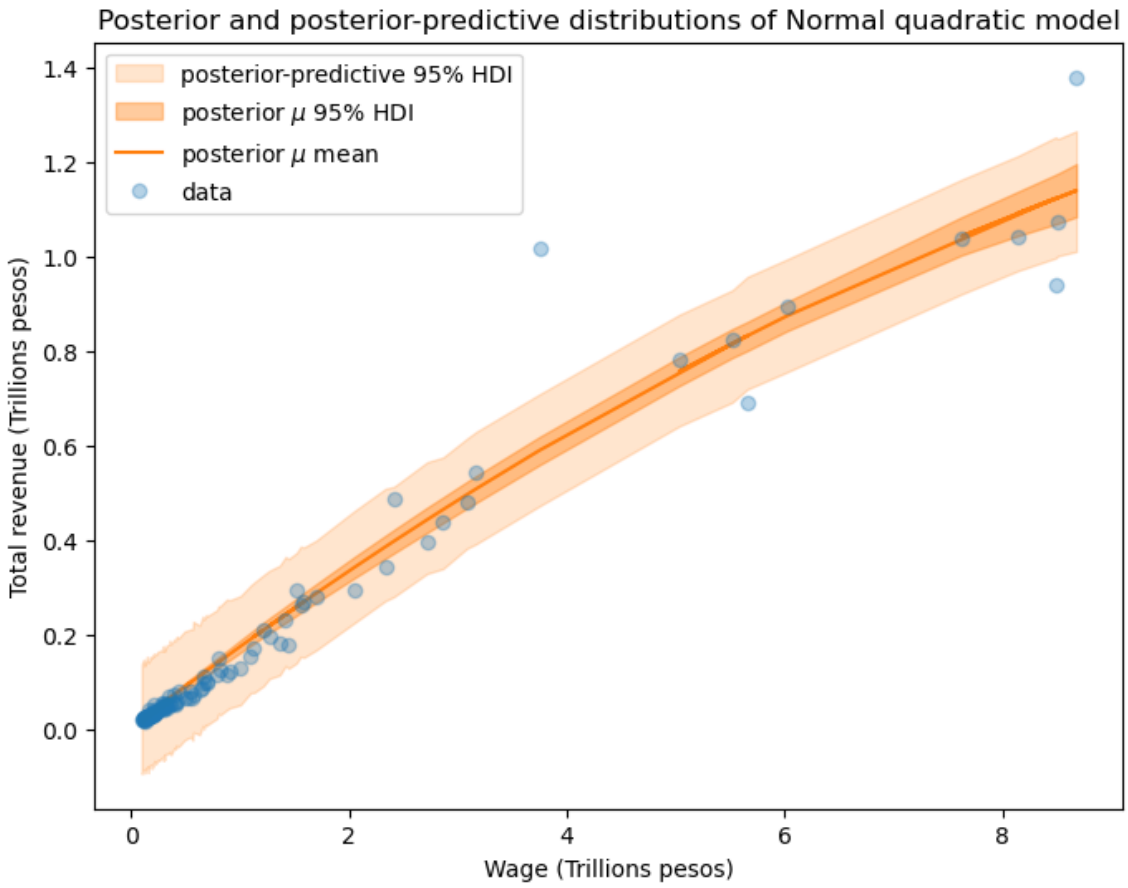
3. **Generating Simulated Heights:**

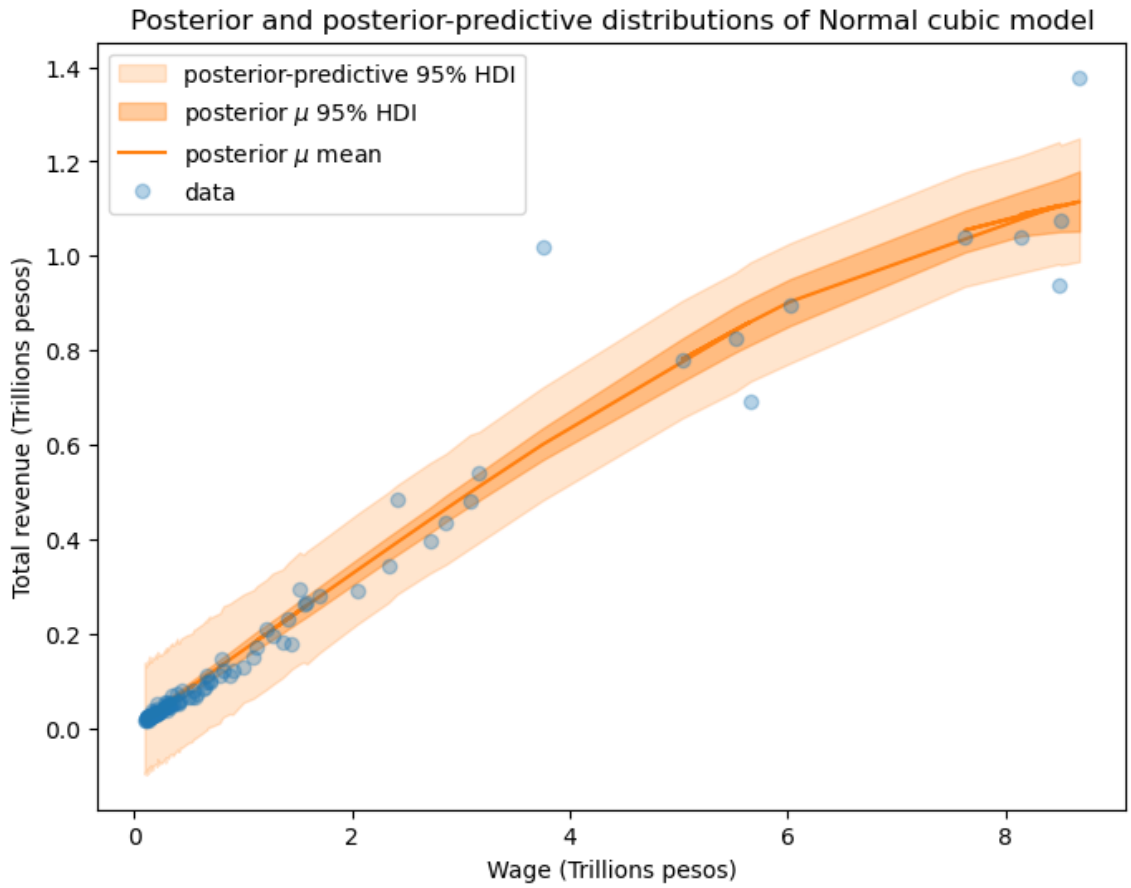
By performing this sampling process for every sample from the posterior distributions of c_0, c_1, and σ, for every value of interest x_i we have a collection of simulated/ predictive observations. Each simulated wage expense in the collection of a specific x_i embodies the uncertainty in both the estimated mean (due to uncertainty in c_0 and c_1) and the observed data's variability (due to uncertainty in σ) -> We get the **posterior-predictive credible interval**

Interpret posterior preditive plots

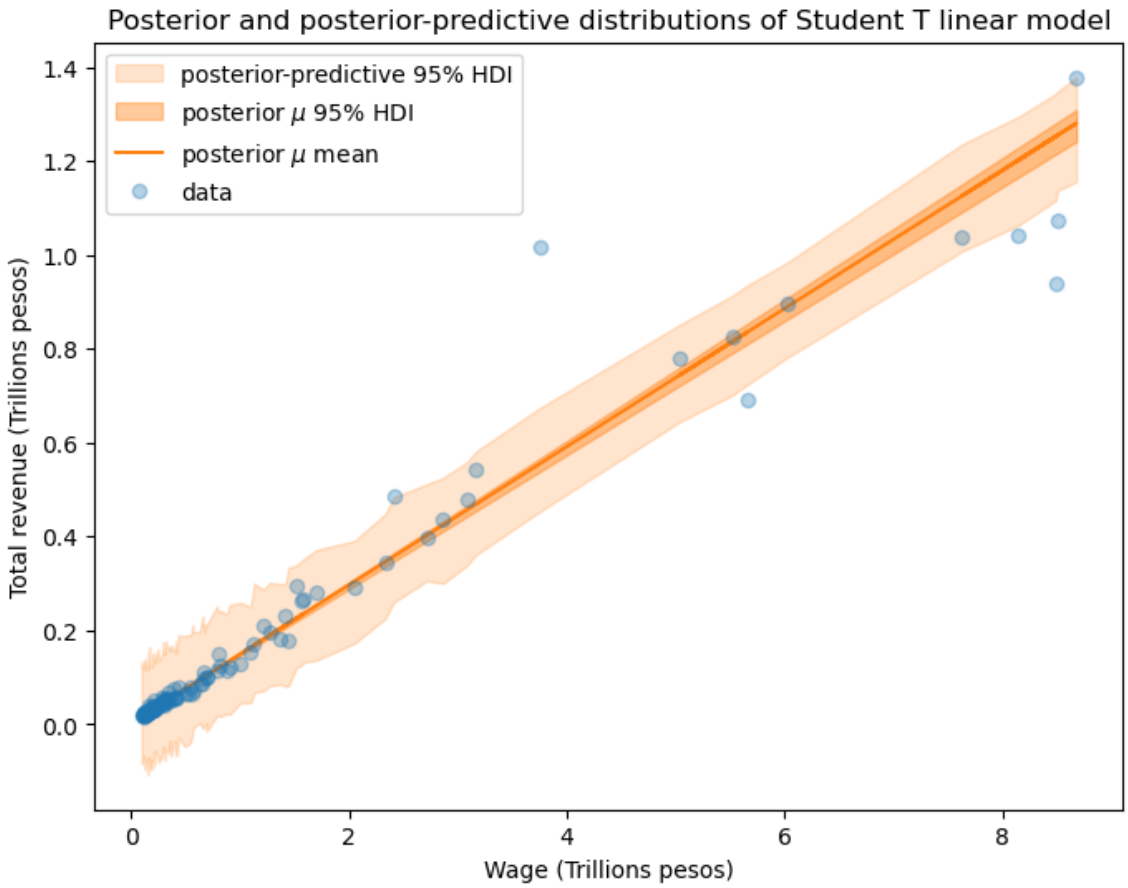
Normal likelihood models: The 1 degree normal linear model shows the sign of shifting downward the regression line to fit the lower right cluster of datapoints (where x values are larger than 8), which is showned by the fact that the regression line lies under most of the datapoints in the range of 2-4 trillion pesos. By incorporating the 2nd and 3rd degree polynomial model, we see that the curve line slightly fit the data better by allowing the regression line to be more fit to both the [2,4] range while still shifting towards the lower right cluster.

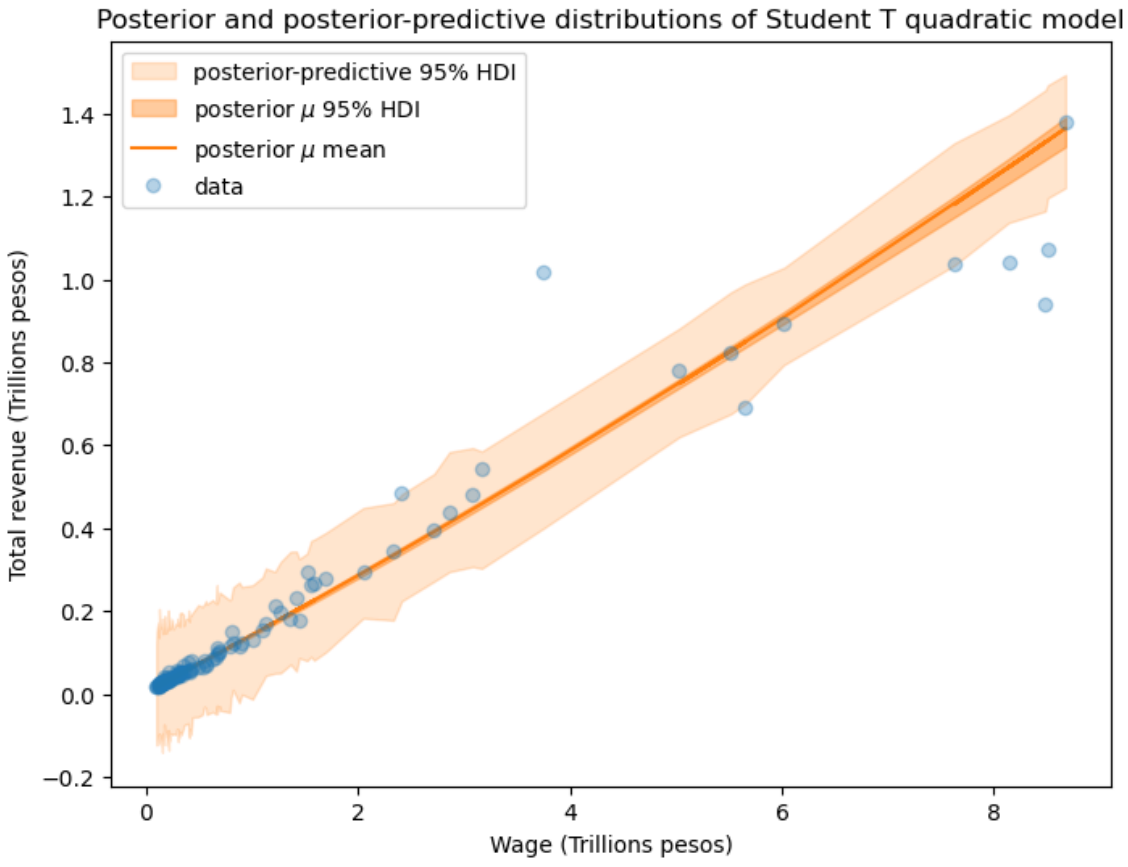

Posterior and posterior-predictive distributions of Normal Linear model
posterior-predictive 95% HDI
posterior μ 95% HDI
posterior μ mean
data
Total revenue (Trillions pesos)
Wage (Trillions pesos)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
2
4
6
8

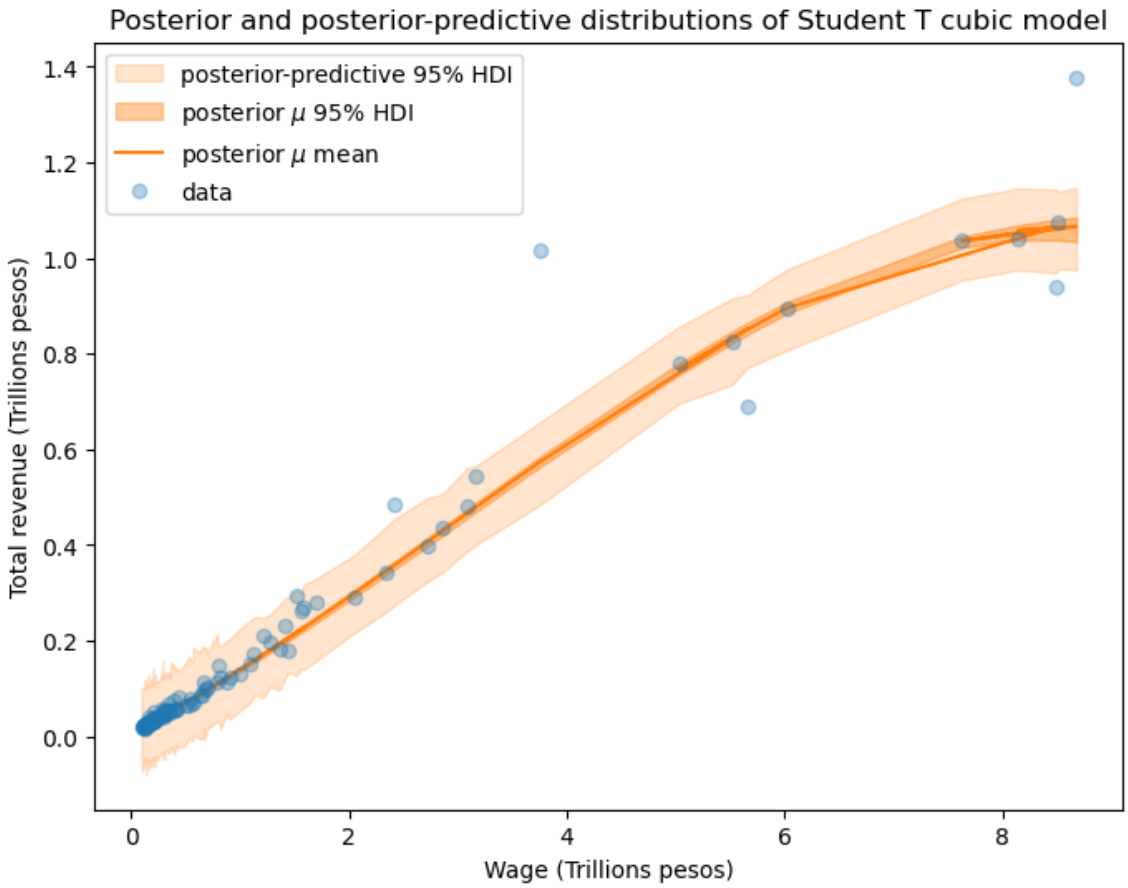

Posterior and posterior-predictive distributions of Normal quadratic model
posterior-predictive 95% HDI
posterior μ 95% HDI
posterior μ mean
data
Total revenue (Trillions pesos)
Wage (Trillions pesos)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
2
4
6
8



Student T likelihood models: For the 1st degree linear model, we can see that the slope of the regression line has increased in comparison to the normal likelihood models, which showcases that the model is more resilient towards the lower right cluster. The quadratic and cubic models show a better fit by fitting the curve line ony through major points.


Posterior and posterior-predictive distributions of Student T linear model
posterior-predictive 95% HDI
posterior μ 95% HDI
posterior μ mean
data
Total revenue (Trillions pesos)
Wage (Trillions pesos)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
2
4
6
8


Posterior and posterior-predictive distributions of Student T quadratic model
posterior-predictive 95% HDI
posterior μ 95% HDI
posterior μ mean
data
Total revenue (Trillions pesos)
Wage (Trillions pesos)
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
2
4
6
8



Let's test these visual interpretation by model comparison using PSIS_LOO.

MODEL COMPARISON

PSIS-LOO (Pareto Smoothed Importance Sampling Leave-One-Out) cross-validation is a method used for assessing model predictive accuracy. The metric ELPD (Expected Log Posterior Predictive Density) is calculated as the sum of the log probabilities of each observation under the model's predictive distribution, excluding the left-out observation. Mathematically, the ELPD is calculated as:

$$ELPD = \sum_{i=1}^{N} \frac{1}{S} \sum_{s=1}^{S} \log p(y_i \mid \theta_{-i,s})$$

where:

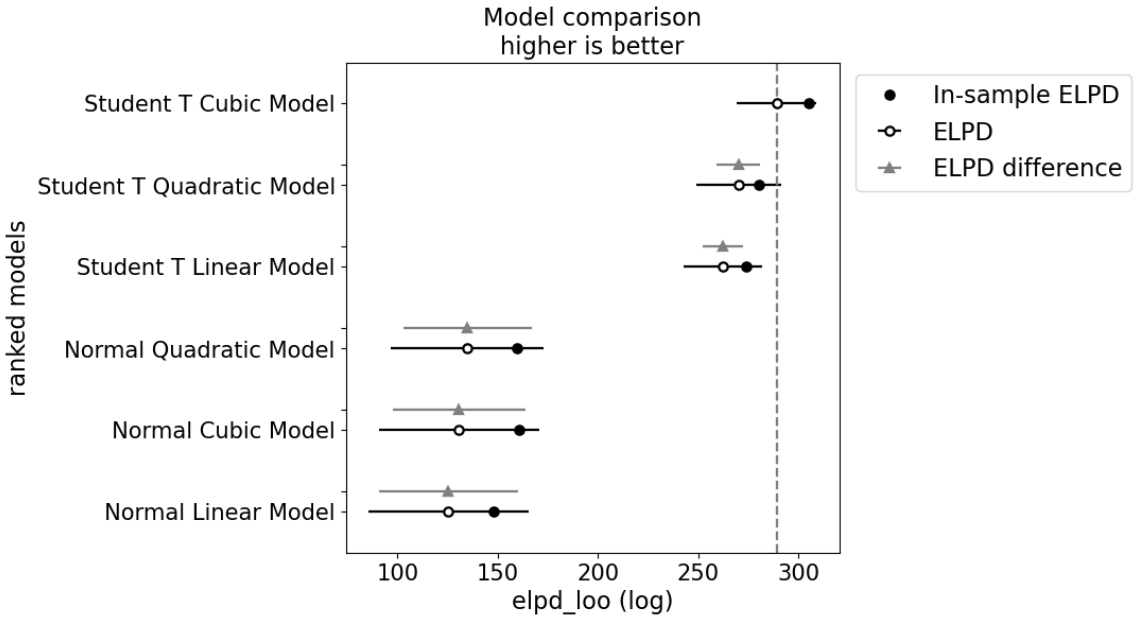
- N is the total number of observations.
- S is the total number of samples
- $logp(y_i \mid \text{data}_{-i})$ is the log of the predictive probability density for $y_i, given the posterior distribution estimated from the remaining n-1 data points.

Overall, ELPD is average log-likelihood of all samples. Higher value of ELPD indicates larger average accuracy, and thus, better ability to fit new, unseen data.

From the plot below, the **Student T Cubic Model** is the most effective, achieving the highest ELPD of **289.28** with a weight of **0.80**. In contrast, the **Normal Linear Model** exhibits the lowest ELPD of **125.26** and a negligible weight, highlighting its inadequacy for this dataset.

- The ELPD difference between the Student T Cubic model and the 2nd rank Student T Quadratic model is 19.114319, which mean that it is $e^{19.114319}$ = 200 million times more likely that the observed data is under the Student T Cubic Model compared to the Student T Quadratic Model. Visually, the ELPD difference error bars barely overlap with the best ELP obtained by the Student T Cubic model.
- All of Student T models are better than Normal likelihood model with a very large ELPD difference shows that the Student T models better accommodate outliers than the Normal Likelihood models (and the outlier influence on the dataset is significant).

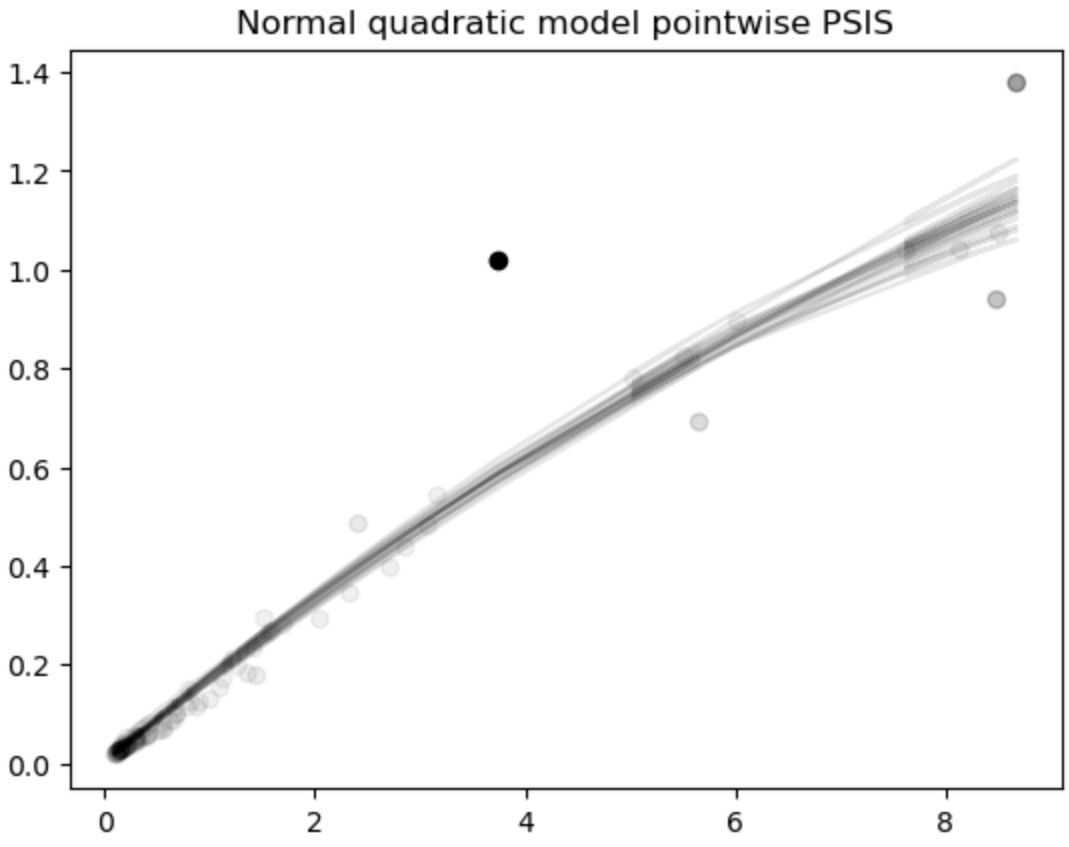
	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
Student T Cubic Model	0	289.283942	15.890006	0.000000	8.019814e-01	19.875188	0.000000	False	log
Student T Quadratic Model	1	270.169622	10.189977	19.114319	1.552471e-01	21.315887	10.737382	False	log
Student T Linear Model	2	262.391333	11.692182	26.892609	0.000000e+00	19.767382	9.875898	True	log
Normal Quadratic Model	3	134.871541	24.768056	154.412400	4.277152e-02	38.049348	32.114909	True	log
Normal Cubic Model	4	130.646143	29.983354	158.637799	0.000000e+00	39.821883	33.118098	True	log
Normal Linear Model	5	125.259327	22.550766	164.024615	6.313172e-12	39.966893	34.720120	True	log

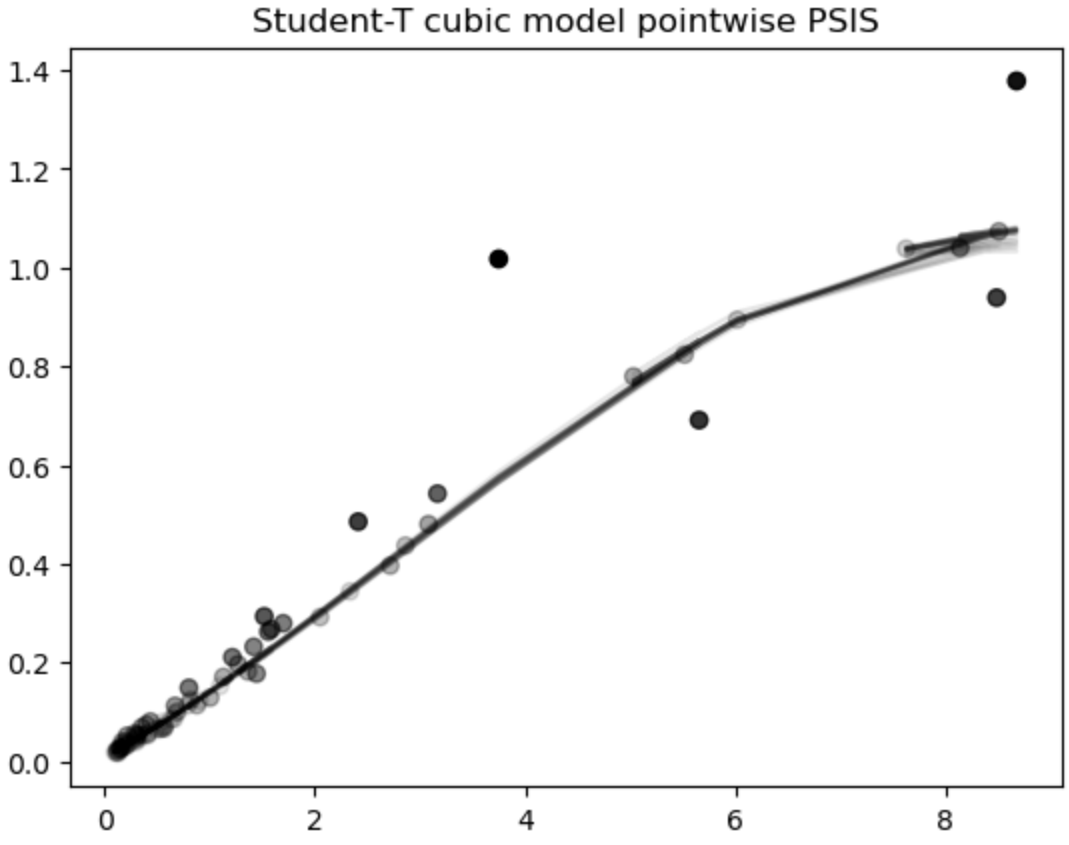


Point wise outlier detection

The PSIS (Pareto Smoothed Importance Sampling) weights indicate the importance of each data point in the model. Higher PSIS weights suggest that a particular observation has a greater influence on the model's predictions and overall performance. In the scatter plots, points that are more opaque (i.e., with higher alpha values) correspond to higher PSIS weights, meaning they contribute more significantly to the model's fit.

-> the Student-T cubic model shows more opaque points, which means it is more robust in handling diverse data distributions.





AI tools: I first use ChatGPT by pasting the assignment prompts and generate a checklist of tasks I need to do. I use it double check and adapt the code from the class.

Reference

Code for Model comparison was adapted from CS 146 session 12, PCW answer key

Code for the Normal, T models, and sampling diagnostics was adapted from CS146 Session 8- Robust Linear Regression pre-class workbook.

Code for the Heteroscedastic models was adapted from CS146 Session 9 - Linear Regression for non-linear data in-class workbook.

Code for plotting posterior-predictive credible intervals and posterior credible intervals for the mean was adapted from CS146 Session 7 - Linear Regression pre-class workbook

In [ ]: