CS146 - Computational Methods for Bayesian Statistics

Linear regression and model comparison (LBA)
Minerva University
Fall 2025
Scheffler, MWh@3PM UTC

November 1, 2025

Contents

1 Introduction

With an extraordinarily high Gini-coefficient of 0.42, Argentina ranks among the top 10% of the most income-unequal countries in the world (Source). This gap appears with particular weight between metropolitan regions like Buenos Aires, Cordoba, or Rosalia, and rural areas. Among many other factors, this manifests in differences in educational attainment between students and schools with different socioeconomic statuses, likely through a mechanism of underfunding, limited access to learning technology and support, and parental support. This analysis will compare different linear models of government data on socioeconomic status and educational performance across Argentinian schools.

2 Data

2.1 Data Source

The data was downloaded from the official Argentinian government website and slightly transformed to allow for a meaningful regression model.

Column AKU of the original dataset `mdesemp_Satisfactorio` shows the number of students per school who achieved 'satisfactory' scores for their age group. The other categories here are "insufficient", "basic" and "advanced", but the advanced category contains no entries. For this analysis, the unit of analysis is each school, so each data point represents a single school captured in the data collection process.

The Secretary of Educational Evaluation (Secretario de Evaluación Educativa) determines the socioeconomic status of a student through a weighted combination of several factors: (1) The parent's educational level, 'overcrowding' in the house (how many people per room), (3) an indicator variable whether the universal child allowance is received (a monthly sum paid to children of unemployed or informally employed parents), and (4) the ownership of computer equipment at home. This way, students are categorized into one of five "Quantiles". The data was transformed to create a new variable that captures the proportion of students in each school that belong to the lowest two categories. Figure 1 shows a scatterplot of the low-income and the math-score variables against each other.

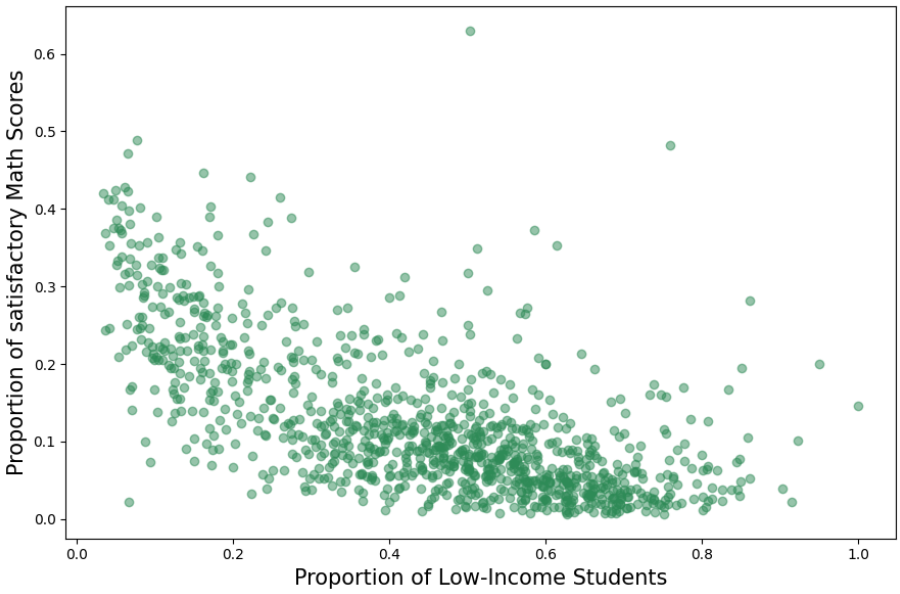

Figure 1: Scatterplot of the modified data.

2.2 Variables

The predictor variable is the proportion of students in a school with low socio-economic status, which is defined as the proportion of students belonging to one of the lowest two groups in the NSE index. This data was standardized to reduce the level of correlation between the higher order terms, which can make

it harder for the sampler to converge. Additionally, this also gives us a more interesting interpretation for the intercept a: rather than showing us the proportion of students with satisfactory math scores in a school without any low-income students, which is highly unrealistic, it now represents the math score for schools with an average low-income population (where $x_s = 0$). We standardize as follows

$$x_{s_i} = \frac{x_i - \bar{x}}{\sigma_x},$$

which converts the units of the x-axis to standard deviations σ. This means that the support, in theory, is infinitely large (however, in practice, it is between -1.98 and 2.72 for this dataset).

The outcome variable is the proportion of math scores above the "satisfactory" threshold, which is the fraction of "basic" and "satisfactory" results to the total number of students. This continuous variable is technically restricted to (0,1) as proportions can neither go below 0 nor exceed 1.

3 Model 1: Polynomial Regression with Normal Likelihood

First, the curved shape of the data in Figure 1 implies that a polynomial linear regression might be an appropriate choice here. Our choice of priors is informed by this standardized model.

Prior for a

The intercept a represents the predicted score for a school with an average low-income population. Since our outcome is a proportion, this parameter must be bounded $[0, 1]$. We use a **Beta distribution** as it naturally enforces this constraint. A $a \sim \text{Beta}(2, 2)$ was chosen, which is a weakly informative prior (over the bounded interval) centered at 0.5 and is symmetric. This encodes our lack of strong prior knowledge about the average score, while still correctly constraining the parameter to the '(0, 1)' interval.

Prior for σ

The residual noise σ must also be small for data bounded on a [0, 1] interval. A $\sigma \sim \text{Beta}(1, 1)$ prior was chosen, which is equivalent to a $\text{Uniform}(0, 1)$. This is a weakly informative prior that enforces the physical constraint that the noise cannot be larger than the entire data range. As we will be using a Normal likelihood later on, values outside the data range might become possible but less likely than for larger prior values.

Priors for b_j (Coefficients)

Because the model is standardized, the polynomial terms can become very large (e.g., if $x_s = 2$, then $x_s^6 = 64$). Experimenting with the look of the prior predictive plots for different degrees, the priors of the weights for the higher-order terms should be increasingly smaller. This encodes our previous knowledge about the relationship between low-income and educaiton outcomes in two ways:

1. **Centering:** We apply a small negative bias to the linear term b_1 based on this "domain knowledge", but we are skeptical of all higher-order terms and center them at 0.
2. **Skeptical Variance:** We are progressively more skeptical of higher-order terms as these can quickly lead to very unrealistic values given our dependent variable. The standard deviation $\sigma_{b,j}$ must shrink rapidly as the degree j increases to reduce the large effect of the x_s^j terms.

After tuning these priors using prior predictive checks on the correct standardized interval (which is reduced to [-2, 2] in the plots as this captures all given data), the following final priors were selected:

$$\begin{aligned}
b_1 &\sim \text{Normal}(\mu = -0.05, \sigma^2 = 0.1^2) \\
b_2 &\sim \text{Normal}(\mu = 0, \sigma^2 = 0.05^2) \\
b_3 &\sim \text{Normal}(\mu = 0, \sigma^2 = 0.01^2) \\
b_4 &\sim \text{Normal}(\mu = 0, \sigma^2 = 0.005^2) \\
b_5 &\sim \text{Normal}(\mu = 0, \sigma^2 = 0.001^2) \\
b_6 &\sim \text{Normal}(\mu = 0, \sigma^2 = 0.0005^2)
\end{aligned} \tag{1}$$

This strategy successfully regularizes the models, preventing the increasingly extreme polynomial wiggles and allowing for a fair comparison.

3.0.1 Model Definition

Finally, a Normal likelihood was chosen as a simple baseline. This model assumes that deviations from the mean are normally distributed, which is a common starting point and models a variety of real-world phenomena. This already raises a limitation of this simple model, because the support of the normal distribution is $(-\infty, \infty)$. This means that this model could, in theory, predict values like 150% for the math scores. These, however, are extremely unlikely under a Normal likelihood. Therefore, the model is defined as follows, with the number of b_j priors equal to the degree d of the model.

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) && \text{(Likelihood)} \\ \mu_i &= a + \sum_{j=1}^{d} b_j \cdot x_{s,i}^j && \text{(Linear Model, Degree } d\text{)} \\ a &\sim \text{Beta}(2, 2) && \text{(Prior for intercept)} \\ b_j &\sim \text{Normal}(\mu_j, \sigma_{b,j}^2) && \text{(Shrinkage priors for } j = 1, \ldots, d\text{)} \\ \sigma &\sim \text{Beta}(1, 1) && \text{(Prior for residual std. dev.)} \end{aligned} \tag{2}$$

Plotting the prior predictive distribution for all six models shows us that while certainly not all lines are entirely plausible, especially for higher-degree models, none of the lines have very extreme values. Each line on this plot represents a sample from the mean line given the priors for a and the corresponding number of bs.

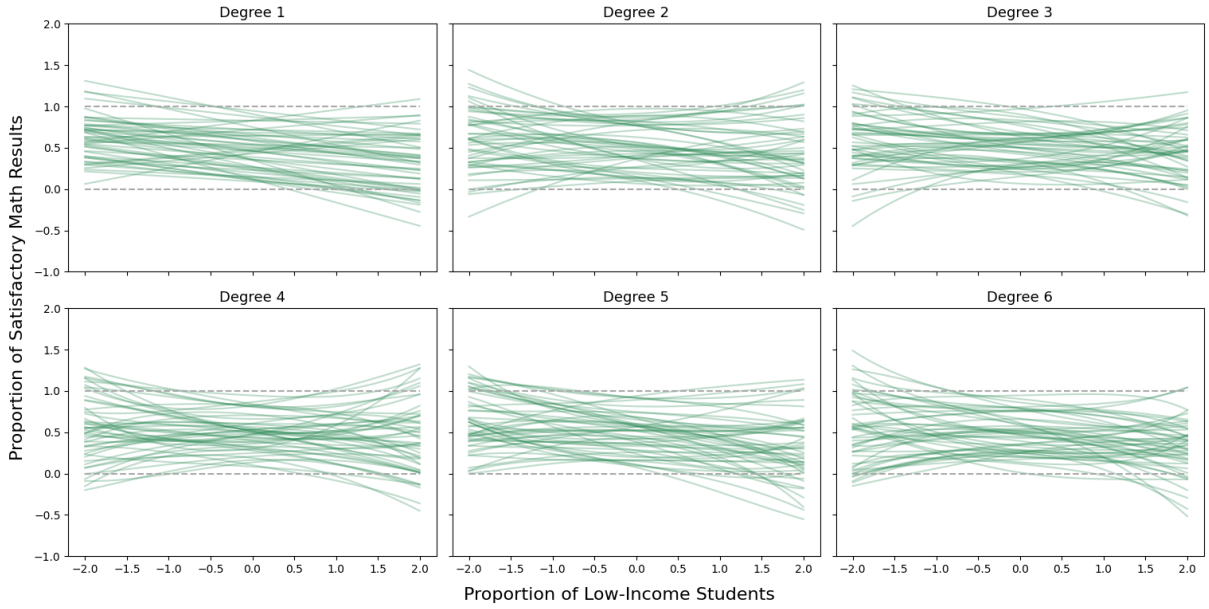

Figure 2: Prior predictive plot for all six models. 50 lines were plotted for every model. Horizontal lines at 0 and 1 were plotted to show that the vast majority of these lines are plausible.

3.1 Posterior Parameter Analysis

Before the models can be compared, we need to make sure that the sampler produced reliable and well-mixed samples from the posterior distribution. For this, a detailed interpretation of the posterior parameter summaries for each of the six polynomial models fit with a Normal likelihood is provided, which, in short, shows that **all samplers work extremely well.**

Model 1 ($d = 1$)

Sampler Diagnostics: The sampler worked perfectly here. The $\hat{R}$ (r_hat) is 1.0 for all parameters. Broadly speaking, this metric compares the variance of all the samples from all chains to the variance of

Table 1: Model 1 Posterior Parameter Summary

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a	0.125	0.002	0.120	0.129	0.0	0.0	4432.0	2793.0	1.0
b1	-0.064	0.002	-0.068	-0.060	0.0	0.0	3974.0	3077.0	1.0
sigma	0.072	0.002	0.069	0.075	0.0	0.0	4335.0	2861.0	1.0

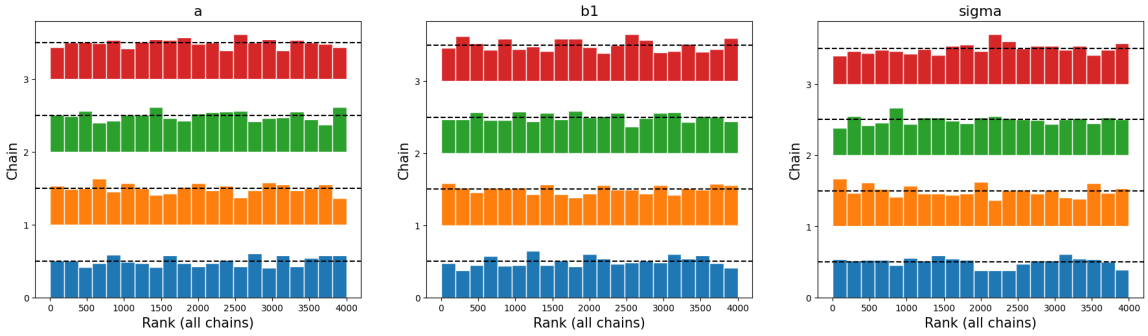

Figure 3: Rank Plots for the three parameters in Model 1. 4 chains were run with 2000 draws each.

each individual chain. The more similar these two are, the close the $\hat{r}$ is to 1.0 and implies that all chains explored the entire posterior space. Next, obtaining at least a few hundred samples in the `ess_bulk` for each parameter is a common threshold to have large enough and less auto-correlated samples that are useful for the compuation of the posteriors. The 'ess_bulk' and 'ess_tail' values are all very high (well over 1000), indicating the chains are well-mixed and the estimates are reliable. Additionally, the rank plots show that each chain explored the entire posterior space: despite small fluctuations, the distributions look approximately uniform and there are no extreme peaks or 'gaps' which would indicate that the sampler got stuck.

Parameter Interpretation:

- **a = 0.125:** The intercept shows that for a school with an *average* low-income population, the predicted proportion of students with satisfactory math scores is 12.5%. The 94% HDI is extremely narrow [12.0%, 12.9%], showing the model is very confident in this estimate.
- **b1 = -0.064:** This coefficient is clearly negative, with an HDI of [-0.068, -0.060] that is far from zero. This suggests a strong, linear negative relationship: for every one standard deviation increase in the low-income population, math scores are predicted to decrease by 6.4 percentage points.
- **sigma = 0.072:** This represents the estimated standard deviation of the "noise" or unexplained variance.

Model 2 ($d = 2$)

Table 2: Model 2 Posterior Parameter Summary

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a	0.101	0.003	0.096	0.106	0.0	0.0	3003.0	2941.0	1.0
b1	-0.060	0.002	-0.064	-0.056	0.0	0.0	3316.0	2313.0	1.0
b2	0.023	0.002	0.020	0.027	0.0	0.0	2928.0	2632.0	1.0
sigma	0.067	0.001	0.065	0.070	0.0	0.0	3566.0	2828.0	1.0

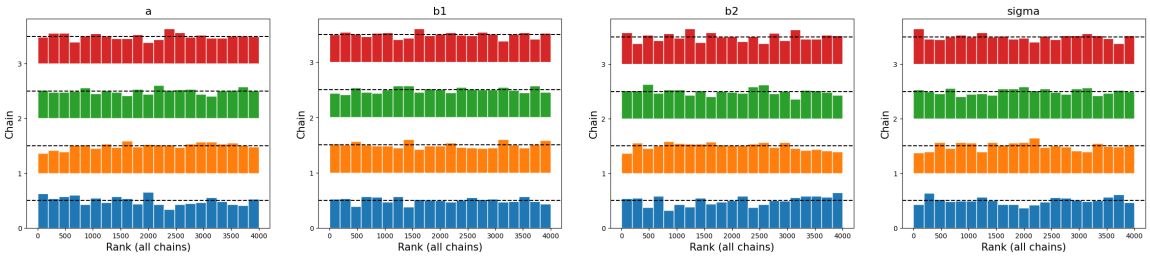

Figure 4: Rank Plots for the three parameters in Model 2. 4 chains were run with 2000 draws each.

Sampler Diagnostics: The sampler worked perfectly again. All $\hat{R}$ values are 1.0 and all ESS values are high, indicating reliable estimates. The rank plots are roughly uniformly distributed without any gaps or major peaks.

Parameter Interpretation:

- **a = 0.101:** The intercept is now 10.1%, slightly lower than in Model 1, as its meaning is now the predicted score at the average, given a more U-shaped curve.
- **b1 = -0.060:** The linear term remains strongly negative.
- **b2 = 0.023:** This is the new parameter. The 94% HDI is [0.020, 0.027], which is clearly and confidently positive. This $b_2 > 0$ is what creates the "U-shape" (a convex curve) that we saw in the data. This might be evidence that a quadratic model is a better fit than a simple linear one.
- **sigma = 0.067:** The residual noise has decreased from 0.072 to 0.068, confirming that this model is explaining more of the variance than Model 1.

Model 3 ($d = 3$)

Table 3: Model 3 Posterior Parameter Summary

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a	0.100	0.003	0.095	0.105	0.0	0.0	2407.0	2778.0	1.0
b1	-0.055	0.004	-0.063	-0.046	0.0	0.0	2146.0	2003.0	1.0
b2	0.024	0.002	0.020	0.027	0.0	0.0	2580.0	2616.0	1.0
b3	-0.002	0.002	-0.005	0.001	0.0	0.0	2322.0	2471.0	1.0
sigma	0.068	0.001	0.065	0.070	0.0	0.0	4024.0	2965.0	1.0

Sampler Diagnostics: The diagnostics are still good, but note the 'ess_bulk' values are starting to drop (e.g., 1800). This indicates the sampler is working harder to explore the parameter space, a sign of increasing model complexity and parameter correlation (multicollinearity).

Parameter Interpretation:

- **b2 = 0.024:** The quadratic term remains strong and positive, with an identical estimate to Model 2.

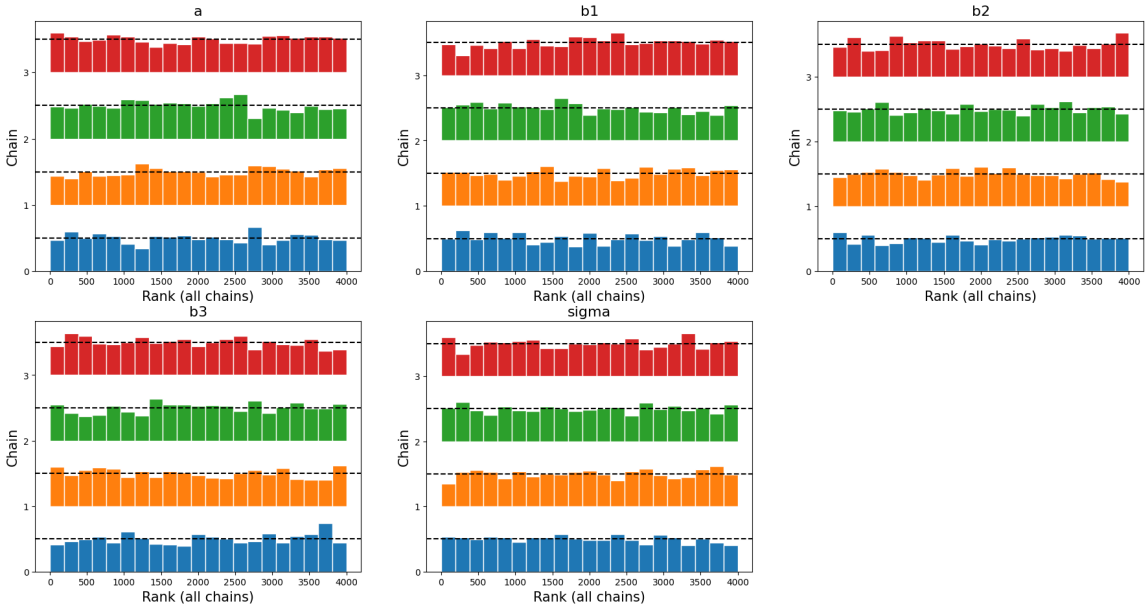

Figure 5: Rank Plots for the three parameters in Model 3. 4 chains were run with 2000 draws each.

- **b3 = -0.002:** This is the key parameter. The 94% HDI is [-0.005, 0.001] which is very close to zero and means that this parameter is not providing a lot of new information; its effect is not statistically different from zero. This is strong evidence that Model 3 will be more complex albeit not a better 'fit' compared to Model 2.
- **sigma = 0.068:** The residual noise is almost unchanged from Model 2, suggesting that adding the b_3 term does not strongly improve the model.

Model 4 ($d = 4$)

Table 4: Model Parameter Summary

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a	0.106	0.003	0.100	0.112	0.0	0.0	2356.0	2684.0	1.0
b1	-0.050	0.004	-0.059	-0.042	0.0	0.0	2297.0	2382.0	1.0
b2	0.007	0.005	-0.002	0.016	0.0	0.0	1821.0	2188.0	1.0
b3	-0.005	0.002	-0.008	-0.001	0.0	0.0	2283.0	2118.0	1.0
b4	0.005	0.001	0.002	0.007	0.0	0.0	1866.0	2317.0	1.0
sigma	0.067	0.001	0.064	0.070	0.0	0.0	3482.0	2539.0	1.0

Sampler Diagnostics: All metrics are great, with the ess_bulk values still being slightly lower as for the first two models.

Parameter Interpretation:

- **b2 = 0.007:** This is a strong change from the previous models. The strong quadratic term from Model 2 has now shrunk, with an HDI of [-0.002, 0.016] that now clearly overlaps with zero. Changing coefficients suggest that the posterior distribution tries to fit different 'wiggles' through the data which might imply it is picking up on very localized patterns, potentially showing overfitting.
- **b3 = -0.005** and **b4 = 0.005:** The model is "splitting" the U-shape effect between b_2, b_3, and b_4. These alternating coefficients create more 'wiggles'. The fact that b_2's effect is no longer clearly positive shows that this model is less interpretable and might be fitting noise too closely.

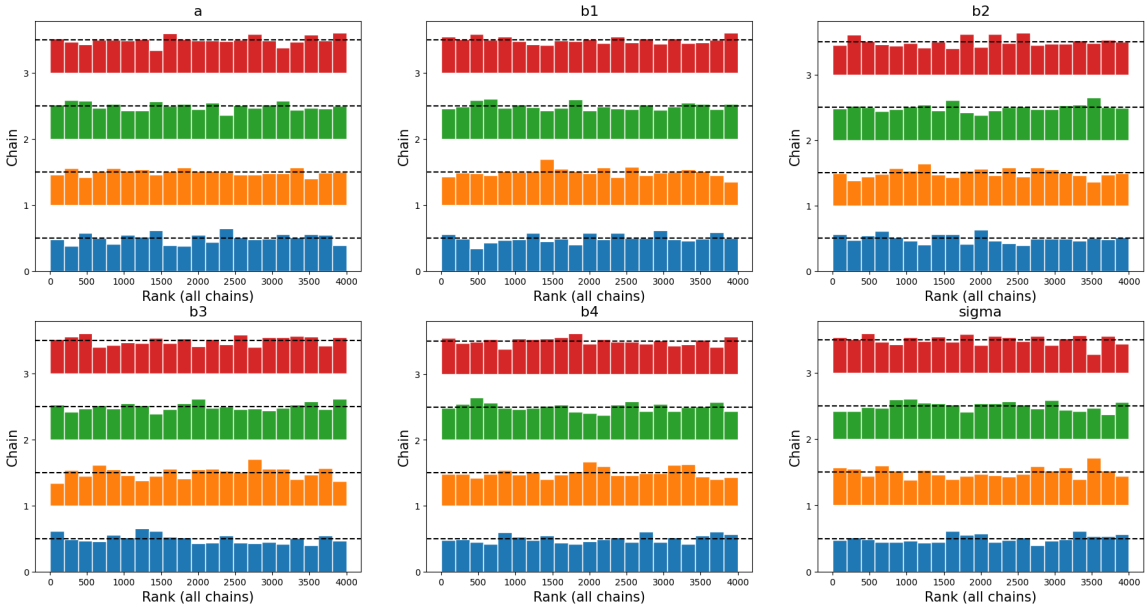

Figure 6: Rank Plots for the three parameters in Model 4. 4 chains were run with 2000 draws each.

Table 5: Model 5 Posterior Parameter Summary

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a	0.106	0.003	0.100	0.112	0.0	0.0	2386.0	2616.0	1.0
b1	-0.051	0.005	-0.061	-0.042	0.0	0.0	2048.0	2413.0	1.0
b2	0.006	0.005	-0.004	0.016	0.0	0.0	1823.0	1908.0	1.0
b3	-0.004	0.004	-0.011	0.003	0.0	0.0	1677.0	2205.0	1.0
b4	0.005	0.002	0.002	0.008	0.0	0.0	1897.0	1856.0	1.0
b5	-0.000	0.001	-0.002	0.001	0.0	0.0	1777.0	2306.0	1.0
sigma	0.067	0.001	0.064	0.070	0.0	0.0	3237.0	2775.0	1.0

Model 5 ($d = 5$)

Sampler Diagnostics: The ‘ess_bulk‘ values are continuing to drop, now as ”low” as 8500-1900. This is a clear sign of increasingly poor sampling efficiency due to an overly complex, overparameterized model.

Parameter Interpretation:

- **b2, b3, b5:** All of these coefficients have HDIs that clearly overlap with zero. This is evidence that these parameters might simply be fitting noise because the shape of the graph strongly changes in each posterior sample. The model has no confidence in their true values.

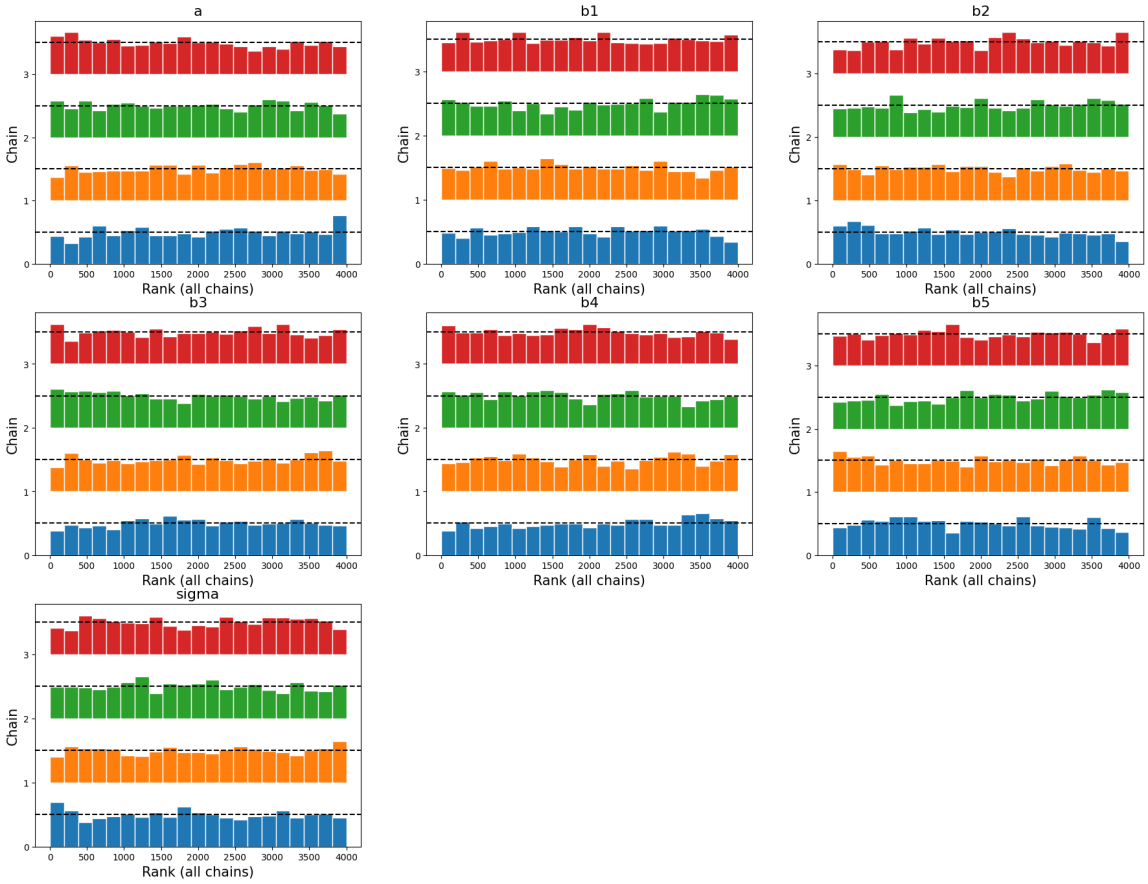

Figure 7: Rank Plots for the three parameters in Model 5. 4 chains were run with 2000 draws each.

Model 6 (Degree 6)

Table 6: Model 6 Posterior Parameter Summary

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a	0.106	0.003	0.100	0.112	0.0	0.0	2574.0	2596.0	1.0
b1	-0.051	0.005	-0.060	-0.041	0.0	0.0	2345.0	2419.0	1.0
b2	0.007	0.006	-0.005	0.018	0.0	0.0	2001.0	2407.0	1.0
b3	-0.003	0.004	-0.011	0.003	0.0	0.0	2076.0	2158.0	1.0
b4	0.005	0.003	0.000	0.010	0.0	0.0	1943.0	2173.0	1.0
b5	-0.000	0.001	-0.002	0.001	0.0	0.0	2264.0	2173.0	1.0
b6	0.000	0.000	-0.001	0.001	0.0	0.0	2259.0	2499.0	1.0
sigma	0.067	0.001	0.064	0.070	0.0	0.0	2955.0	2187.0	1.0

Sampler Diagnostics: While all metrics suggest successful sampling, this model has similar ess_bulk metrics as Model 5, which are lower than for the simpler models.

Parameter Interpretation:

- **b2, b3, b5, b6:** Four of the six slope coefficients have HDIs that cross zero. Similarly to model 5, the model is adding 'wiggles' that might follow the shape of the data too closely.

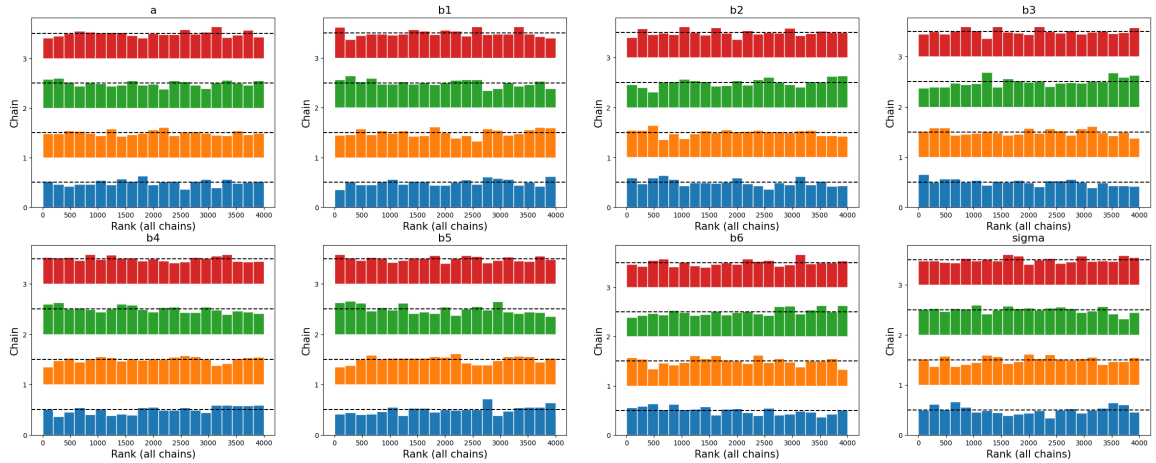

Figure 8: Rank Plots for the three parameters in Model 6. 4 chains were run with 2000 draws each.

3.2 Posterior and Posterior Predictive

Having established that the sampler worked successfully for all six models above, we can plot the posterior and posterior predictive distributions for the models. These confirm our intuition from the posterior results in that from Model 4 onward, the data points at the right end of the distribution more and more strongly affect the shape of the mean in that area. At the same time, the 99% interval around the mean becomes increasingly large in that area, which is plausible as only few data points with a large variance around the mean line determine the shape of the graph here. Further, Models 2 and 3 look very similar to each other,

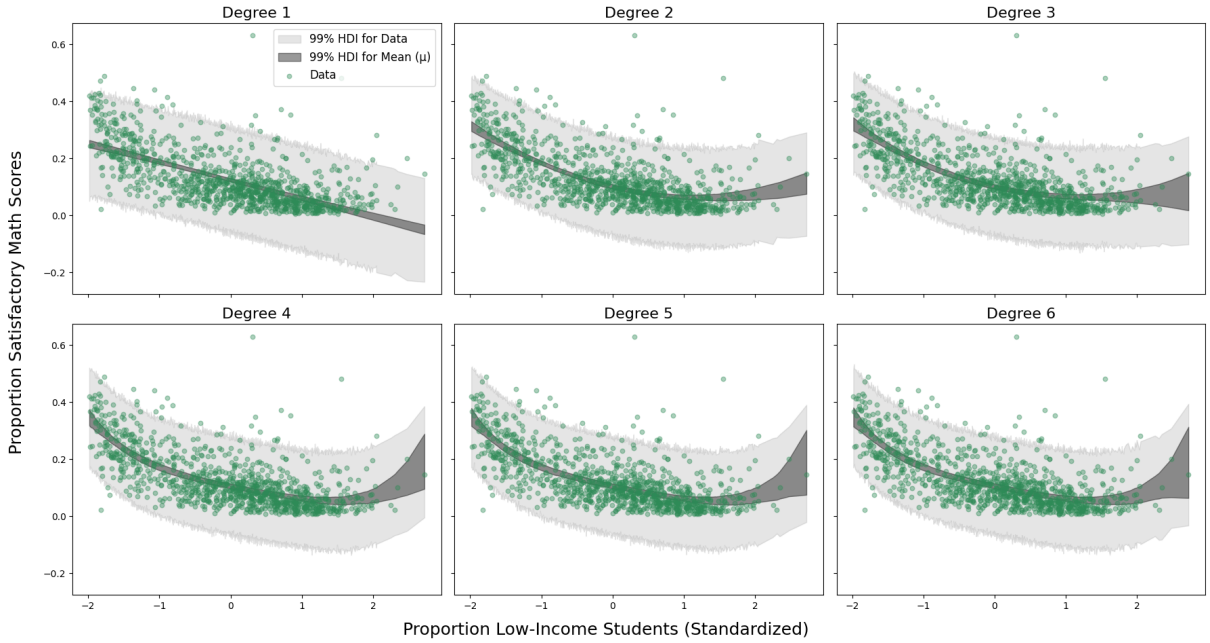

Figure 9: Posterior and posterior predictive plots for all six models showing the 99% interval for μ in the dark shade, and then 99% interval for the posterior predictive in the light shade.

3.3 Polynomial Model Comparison

To determine which of the models has the highest predictive accuracy, we need to compute the expected log-pointwise predictive density (ELPD). Ultimately, this aims to obtain values for the out-of-sample deviance as these can help us compare which models are overfitting to the in-sample data. In even simpler words, this shows how well the model can make predictions for new, unseen data. One method

for this is computing the log-pointwise-predictive density: this is done by considering the log-likelihood of each data point $p(y_i|\theta)$ over *all* posterior draws of the parameters. Finally, the log-estimates for each data point are averaged to obtain the LPPD. However, this is often a poor estimate for unseen data as it is computed using only the known data. Using LOO-CV (Leave-One-Out Cross Validation) is a better approach as it simulates out-of-sample predictions.

For this model comparison, Pareto-Smoothed Importance Sampling (PSIS) was preferred for multiple advantages. Firstly, it adjusts the weights of highly unlikely observations that are far away from the 'bulk' of the other observations, leading to more stable predictions. Additionally, it provides a warning when the parameter k of the Pareto distribution exceeds a certain threshold, implying that importance sampling did not work reliably. Using PSIS-LOO, we can now compare the ELPD for the six models, where higher values show a larger predictive accuracy. Let us plot the resulting ELPD values, as well as their difference, to get an intuition about how these models compare.

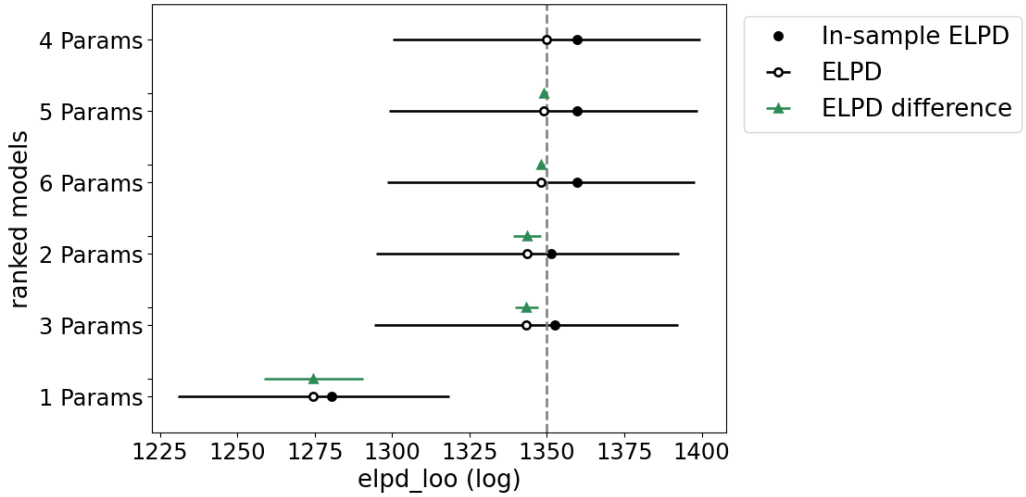

Figure 10: Ranked model comparison using PSIS-LOO. The horizontal lines show the 68% intervals around the mean OOS-ELPD. The green markers and intervals show the difference of all models to the highest-ranked model with the horizontal dashed line being the line of no difference. The "In-sample ELPD" corresponds to LPPD.

Table 7: Model Comparison Results between the six polynomial models (LOO)

Model	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning
4 Params	0	1349.896720	9.805592	0.000000	8.929×10^{-1}	49.810213	0.000000	False
5 Params	1	1349.728696	10.037191	0.168024	0.0	49.694572	0.289137	False
6 Params	2	1349.294061	10.519687	0.602660	0.0	49.767664	0.313272	False
2 Params	3	1343.885629	7.452855	6.011091	6.481×10^{-16}	48.966066	4.292647	False
3 Params	4	1343.565740	8.700646	6.330980	0.0	48.982458	3.650504	False
1 Params	5	1274.700523	5.801444	75.196197	1.071×10^{-1}	43.753945	16.142205	False

While this shows us a preference for the fourth degree polynomial model, the plot also indicates that all models with two more more parameters have very similar ELPD values (e.g. `ELPD.4-ELPD.5`$= 0.896$). Similarly, note that the green horizontal lines on the plot only show the 68% confidence intervals for the difference in ELPD values between the ranked models. However, a 99% confidence interval, which has been used throughout this paper to establish a higher confidence in the difference between the models, is not shown by default. For normally distributed data, which we assume here as the likelihood is normal, a 99% confidence interval corresponds to a z-value of 2.576. Hence, we need to compare whether $\text{elpd\_diff} \pm 2.576 \cdot \text{dse} < 0$.
Table 8 shows that with 99% level of certainty, the difference between models 2, 3, 5, and 6 and the best-ranked model 4 is *not plausibly different*. This means that it is possible for the ELPDs of models 2, 3, 5, and 6 to be greater than the ELPD of model 4. We can only conclude with high confidence that the

Table 8: Model Comparison Data

Model	Lower	Upper	Crosses
5 Params	-0.576793	0.912840	True
6 Params	-0.204328	1.409647	True
2 Params	-5.046767	17.068950	True
3 Params	-3.072717	15.734677	True
1 Params	33.613877	116.778518	False

ELPD of Model 1 is lower than the ELPD of Model 4 at this confidence level. In this case, we should opt for the simplest out of all these models, as we are *not confident* that increasing the degree of the model beyond deg = 2 actually increases the ELPD. If a simpler model with deg = 2 has high predictive accuracy, why make it more complicated?

Therefore, Model 2 is our preferred model among the six polynomial models.

4 Model 2: Polynomial Regression with Student-T Likelihood

While Model 2 'beats' the other polynomial models in its predictive accuracy, some points are very far away and extremely unlikely under its Normal likelihood model assumption. Given the credibility of the source of the dataset, we cannot simply 'remove' these points as they are neither likely to be due to measurement errors nor in any way implausible. While somewhat surprising, there might be schools with high levels of low-income students that still have high average math scores. Hence, we could use the thicker-tailed Student-T distribution to account for the high dispersion of the data points. Under this new model, more extreme data points (especially in the right half of the distribution) are more plausible to be observed, and captured in the posterior predictive interval.

The Student-T distribution is a common addition to the Normal as it includes an additional "degrees of freedom" parameter ν which allows the model to have heavier tails. For this model, we chose the priors based on the Degree 2 model and add the ν parameter.

- **New prior for ν:** We use $\nu \sim \text{HalfNormal}(30)$. This is a common, weakly-informative prior (as $\nu > 0$) that is centered at 0 and allows for a wide range of positive values. In fact, if the data is not heavy-tailed, the posterior for ν will move towards 30 (or higher), at which point the Student-T distribution closely approaches the shape of a Normal distribution.

The final Degree 2 Student-T model is defined as follows:

$$\begin{aligned}
y_i &\sim \text{StudentT}(\nu, \mu_i, \sigma) && \text{(Likelihood)} \\
\mu_i &= a + b_1 \cdot x_i + b_2 \cdot x_i^2 && \text{(Linear Model, Degree 2)} \\
a &\sim \text{Beta}(4, 2) && \text{(Prior for intercept)} \\
b_1 &\sim \text{Normal}(-0.05, 0.2^2) && \text{(Prior for } b_1\text{)} \\
b_2 &\sim \text{Normal}(0, 0.05^2) && \text{(Prior for } b_2\text{)} \\
\sigma &\sim \text{Uniform}(0, 1) && \text{(Prior for residual std. dev.)} \\
\nu &\sim \text{HalfNormal}(30) && \text{(Prior for degrees of freedom)}
\end{aligned} \tag{3}$$

Sampler Check

Before we proceed to the final comparison between the 2-degree Normal and Student-T models, we first need a successful sampling process again.

Sampler Diagnostics: The sampler worked perfectly. All $\hat{R}$ values are 1.0 and all ESS values are very high, indicating the chains are well-mixed and the estimates are highly reliable.

Table 9: Student T Model Posterior Parameter Summary

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a	0.093	0.002	0.088	0.097	0.000	0.000	2949.0	2918.0	1.00
b1	-0.063	0.002	-0.066	-0.059	0.000	0.000	3674.0	2751.0	1.01
b2	0.024	0.002	0.021	0.027	0.000	0.000	2971.0	2424.0	1.00
nu	3.437	0.423	2.688	4.224	0.007	0.007	3255.0	3166.0	1.00
sigma	0.045	0.002	0.042	0.049	0.000	0.000	3121.0	2859.0	1.00

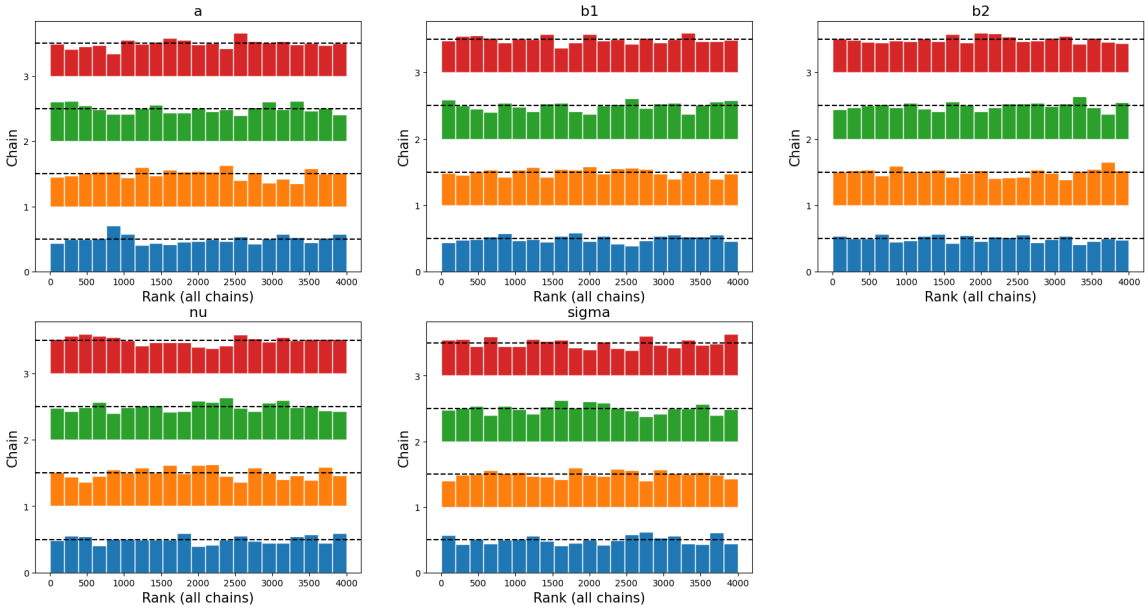

Figure 11: Rank Plots for the three parameters in the Student-T Model. 4 chains were run with 2000 draws each.

Parameter Interpretation:

- **a = 0.093, b1 = -0.063, b2 = 0.024:** The estimates for the polynomial coefficients are very similar to the Normal model (a=0.101, b1=-0.063, b2=0.024). This shows that the main "U-shape" trend is robust. The quadratic term b_2 remains clearly positive, confirming the convex shape.
- **nu = 3.422:** This is the most important parameter. A low 'nu' value, with a 94% HDI of [2.688, 4.224], is definitive proof that the data has heavy tails. The model is extremely confident that a Normal distribution (which is a Student-T with 'nu' $\to \infty$) is a poor fit. This parameter is "soaking up" the effect of the outliers.
- **sigma = 0.045:** In the Normal model, σ was 0.068. Here, it has shrunk to 0.045. This is the core benefit of the Student-T model: by using 'nu' to handle the outliers, σ is now free to model the "true" standard deviation of the bulk of the data, which is much smaller than the Normal model was led to believe.

We can also an interesting relationship between the posterior values of ν and σ. This is plausible as both are ways to make values further away from the mean line more plausible: when σ increases, the distribution is naturally wider which makes more extreme values less surprising. In those cases, the distribution does not need heavy tails because the variation of data points is already 'covered' by the value of σ. Following the same logic, thicker tails (a lower value of ν) do not require a standard deviation as high as the tails already include them, even with a low σ.

4.1 Posterior and Posterior Predictive

Having asserted the success of the sampling process, we can compare the posterior and posterior predictive plots for the two models. As intended, the Student T model has a significantly wider 99% interval for its

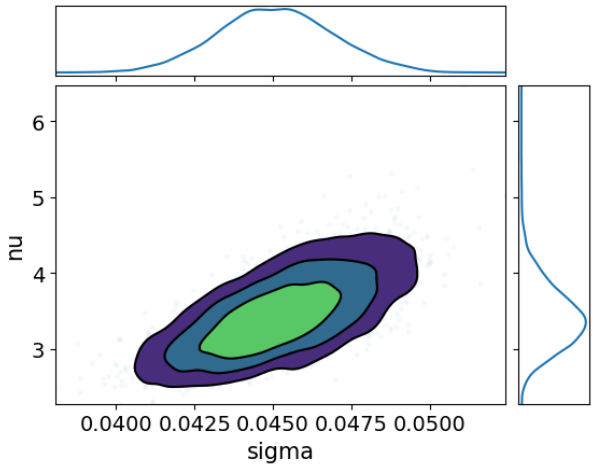

Figure 12: Pair Plot between ν and σ.

posterior predictive. While a few outliers remain, they are expected at this level of confidence (1% of data points are not captured in the posterior predictive interval). At the same time, the 99% interval for the mean has become narrower for the right end of the distribution. While the Normal model was strongly influenced by the very variant data points in that area, the Student T distribution is more robust and assigns them to its heavier tails. This suggests that using the Student T likelihood ultimately achieved higher confidence in the mean line while 'allowing' for larger deviations from the mean to occur.

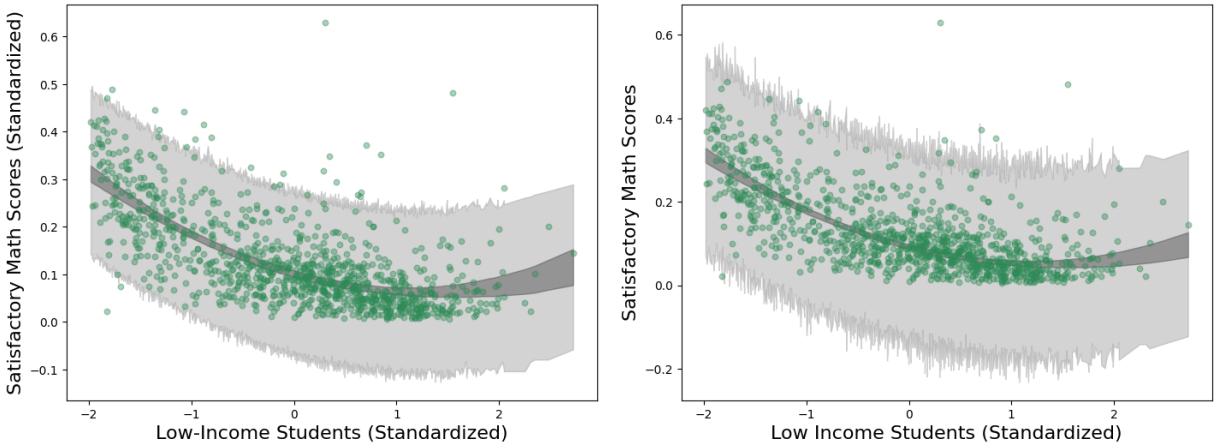

Figure 13: Posterior and Posterior Predictive for the Normal (left) and Student T (right) likelihoods. The dark shaded region is the 99% HDI around μ; the light-shaded area is the 99% HDI of the posterior predictive.

4.2 Final Model Comparison

We can now use the PSIS-LOO again to compare the ELPDs of these two models. Given the discussion on the plots above, it is no surprise that the Student-T model is ranked higher than the quadratic model. This time, the interval around the difference (Figure 6) is also significantly further away from the line of no difference, meaning that we can be 99% confident that the Normal model has less predictive accuracy. Specifically, if we extend the right half of the green line, its standard error, to the right by a factor of $z = 2.576$, it would clearly not cross the dashed line. Mathematically, `elpd_diff`$-2.576\times$`dse`$= 68.185 - 2.576 \times 14.876 = 29.864424 > 0$. At this confidence level, it is not possible that the ELPD of the Normal model is as high as the ELPD for the Student T model. Lastly, the 'weight' column further supports this. It suggests there is a 90.9% probability that the Student-T model will make better predictions on new, unseen data compared to the Normal model.

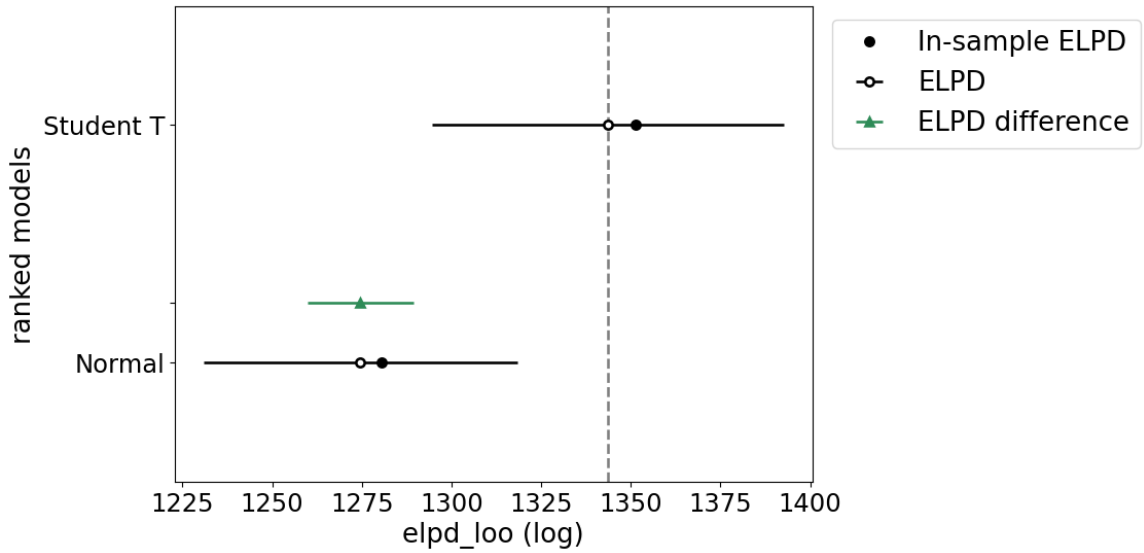

Figure 14: Ranked model comparison between the Normal and the Student-T model using PSIS-LOO. The horizontal lines show the 68% intervals around the mean OOS-ELPD. The green markers and intervals show the difference of all models to the highest-ranked model with the horizontal dashed line being the line of no difference. The "In-sample ELPD" corresponds to LPPD.

Table 10: Model Comparison Results (LOO)

Model	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning
Student T	0	1343.885629	7.452855	0.000000	0.908916	48.966066	0.000000	False
Normal	1	1274.700523	5.801444	69.185106	0.091084	43.753945	14.876572	False

5 Brief Conclusion

This analysis set out to model a linear regression model relationship between the proportion of low-income students and academic performance in Argentinian schools. Our final model, a **Degree 2 polynomial with a robust Student-T likelihood**, nicely models their relationship.

For this, we used a two-stage model comparison: firstly, using PSIS-LOO and the principle of preferring the simplest model if adding higher degrees does not create significant improvements in the ELPD, we determined Model 2 ($d = 2$) to be the model with the highest predictive accuracy.

Secondly, we compared the Normal and Student-T likelihoods as the data is characterized by significant outliers ($\nu \approx 3.4$). The final Student-T model successfully accounted for these outliers, leading to a more robust, confident estimate of the underlying trend.

In real-world terms, our final model (with $a = 0.093$, $b_1 = -0.063$, $b_2 = 0.024$) suggests:

- For an average school, the predicted score is a very low 9.3%.
- The relationship is clearly negative, with scores dropping as the low-income proportion increases from the mean.
- However, the positive quadratic term (b_2) confirms that this negative relationship flattens out but reverses at the extremes. This suggests that the negative impact of a high low-income population "bottoms out" and that some schools with very high proportions of low-income students are performing better than this simple negative trend would predict. A deeper analysis of confounding variables would take this further (e.g. specific funding or talent supporting programs in schools with very high proportions of low-income students).

Ultimately, this demonstrates that robust statistical methods are required to model the relationship between socio-economic status and math scores accurately, preventing the influence of extreme data

points from distorting the relationship between them.

6 References

Kademián, N. (2023, November 28). **Income inequality in the Argentine provinces**. Eurac Research. https://www.eurac.edu/en/blogs/eureka/income-inequality-in-the-argentine-provinces

Secretaría de Educación de la Nación, Subsecretaría de Evaluación e Información Educativa. (2024a). **2011 - 2024 - Diccionario bases aprender anonimizadas**. https://datos.gob.ar/dataset/educacion-base-datos-por-escuela-2023/archivo/educacion_26.8

Secretaría de Educación de la Nación, Subsecretaría de Evaluación e Información Educativa. (2024b). **2024 Base APRENDER - Censal - Secundaria 5-6 año - Agregada - Desempeños de Matematica(in)** [Dataset]. https://ministeriodeeducaciondelanacion-my.sharepoint.com/:x:/g/personal/santiago_pomeranz_educacion_gob_ar/EWg_dZb6tdRKmT4i5vIJG3YBvPrxH8vTMvCMzEo5rlZM1A?e=q6a5jD

The World Bank. (n.d.). **Gini index - Argentina**. Retrieved November 1, 2025, from https://data.worldbank.org/indicator/SI.POV.GINI?most_recent_year_desc=true&year=2023 https://datos.gob.ar/dataset/educacion-base-datos-por-escuela-2023/archivo/educacion_26.8

7 Acknowledgments

Gemini Pro was used solely to quickly convert Python output to tables in LaTeX.

I also want to acknowledge our textbook, Statistical Rethinking, which has been a great resource to consult throughout this assignment.

Using Class Code for this Assignment:

Class Code was taken and adapted from multiple sessions. I put effort into improving visualizations and clarity rather than just copy-pasting code.

- Code from Session 7 (Code 4.39 and 4.40: Prior-predictive plots) was modified to create the prior predictive check (Figure 2).
- Code from Session 8 and my PCW from that session (Code 4.66: Posterior summary) was used to define the polynomial regression models (Figure 9).