Project 1:

Linear Regression and Model Comparison

Minerva University

CS146

Prof. E. Volkan

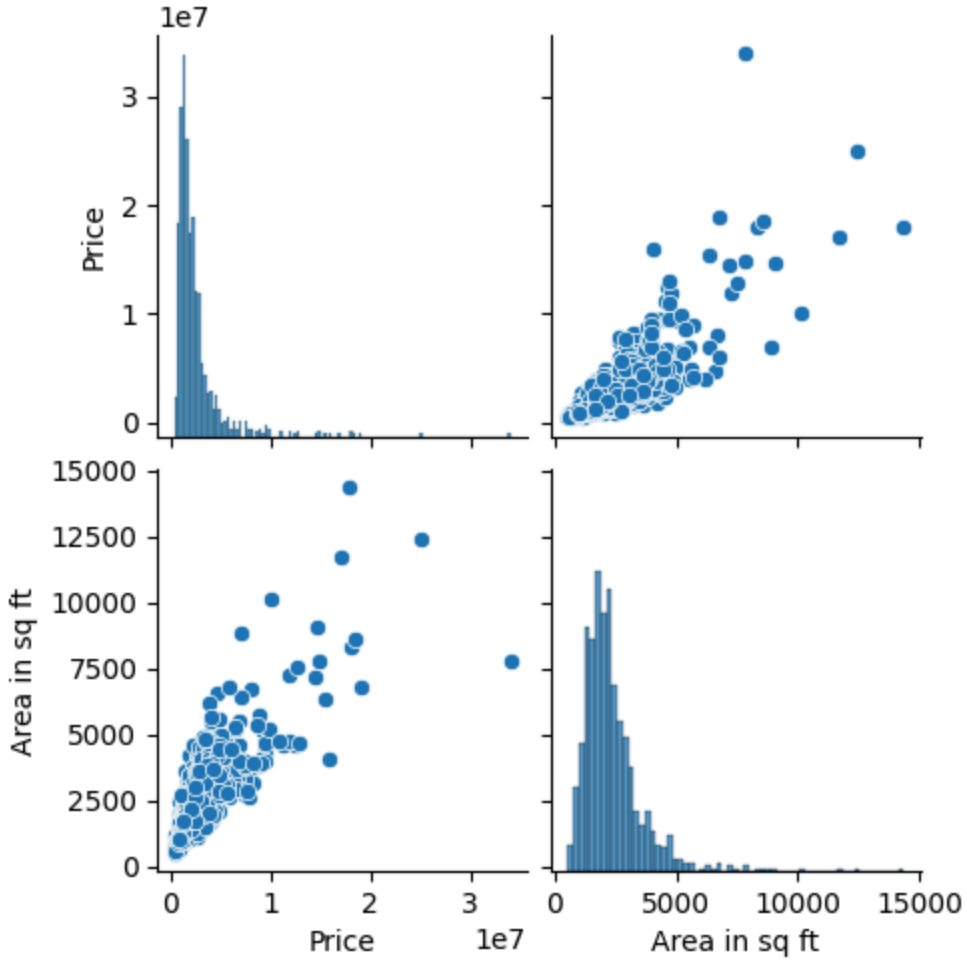
23 February, 2024

1. Introduction

In this project, we aim to develop a model to predict London housing prices using various housing features. Our analysis focuses on the relationship between the size of a house and its price, which is crucial in London due to the limited residential space affecting prices significantly.

Our dataset includes 3,500 data points from Greater London collected by webscrapping, covering different housing types (Kulkarni, 2021). However, to ensure accuracy and relevance, we narrowed our analysis to London proper, specifically to single-family homes labeled "House." This decision led us to exclude data on the number of bedrooms or bathrooms and geospatial data like postal codes. After these adjustments, we have about 1,000 data points for our analysis. Our main question is: How accurately can we predict the prices of single-family homes in London based on their size?

2. Preliminary Data Analysis



Our analysis reveals that the price and area data exhibit a log-normal distribution, with a concentration of data points towards the lower end and fewer instances of higher values. This pattern aligns with the London real estate market, where bigger properties are less common due to constrained real estate space and are more expensive. The scatterplot of area versus price visually supports this observation. Given these characteristics, transforming our data with a logarithmic scale will improve our analysis, making it more suitable for accurately capturing the relationship between house size and price. This is reasonable since both price and area are never negative so this operation can be performed.

The logarithmic transformation applied to our data has effectively linearized the relationship between house size and price. However, the scarcity of data points at the extreme ends of the distribution suggests a higher variance in these areas. Our model must account for this increased variance to ensure accurate predictions across the entire range of property sizes and prices.

3. Methodology

In predicting London housing prices based on house size, we employed linear regression models with distinct likelihood functions, Normal and Student T. The choice of linear regression is

grounded in the assumption that a linear relationship exists between house size and price, a hypothesis supported by preliminary data analysis. The Normal likelihood model is standard for regression analyses, assuming homoscedasticity and that residuals follow a normal distribution. However, the real estate market, particularly in London, is known for its high variability and potential outliers—extraordinarily high or low prices that don't align with the general trend. To accommodate this, we integrated a Student T likelihood model, which is less sensitive to outliers due to its heavier tails, providing a more robust analysis against anomalous data.

4. Model Description

Below we describe the mathematical models for each method we tried and interpretation for the priors of each. These examples are for degree 1 polynomials although higher order polynomials were also used to model the data, with increase in 1 normally distributed coefficient variable for each increase in degree.

Normal Linear Regression

$\alpha \sim Uniform(0,\ 20)$

$\beta \sim Normal(1,\ 5)$

$\mu = \alpha + \beta \cdot area_i$

$\sigma \sim Uniform(0, 50)$

$price_i \sim Normal(\mu,\ \sigma)$

α: Intercept, theoretical starting price for a house with zero area.

β: Slope, change in price for each additional area unit on a logarithmic scale.

$area_i$: Size of the i-th house.

μ: Predicted price for the i-th house.

σ: Variability in price around the predicted value.

$price_i$: Predicted price for the i-th house.

Student - t Linear Regression

$\alpha \sim Uniform(0, 20)$

$\beta \sim Normal(1, 5)$

$\mu = \alpha + \beta \cdot area_i$

$\sigma \sim Uniform(0, 50)$

$\nu \sim HalfNormal(10)$

$price_i \sim StudentT(\mu, \sigma, \nu)$

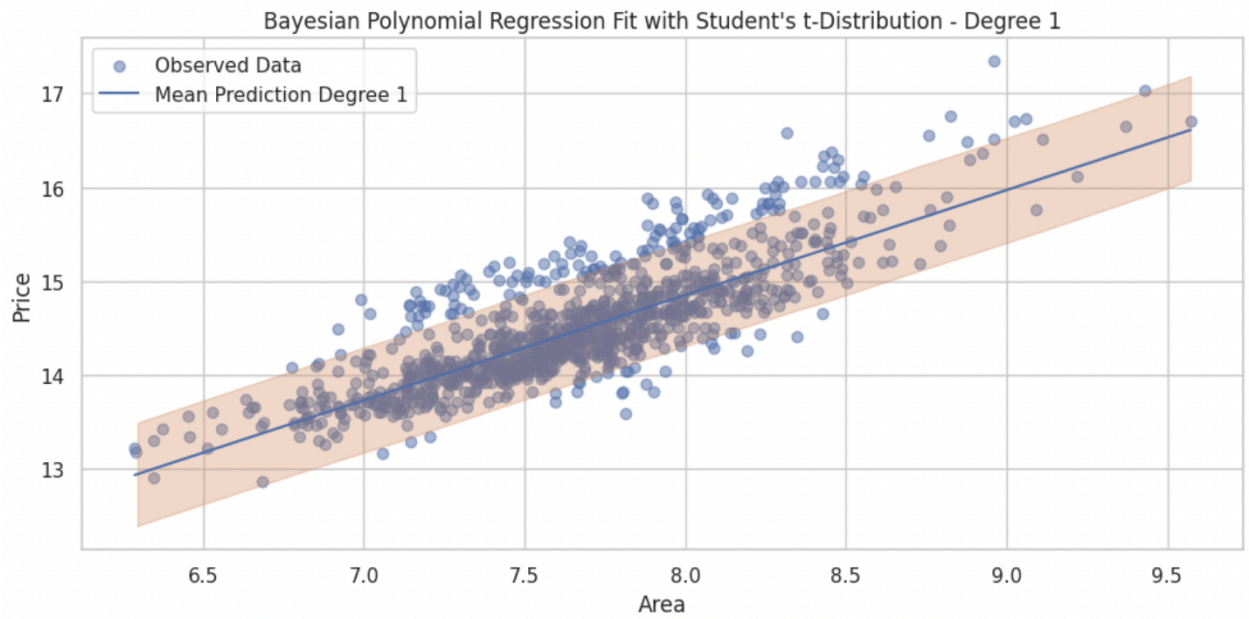
This uses a similar approach with the exception that it uses the student t distribution which differs from the normal distribution in that it has greater mass in the tails and therefore is able to account for outliers in data.

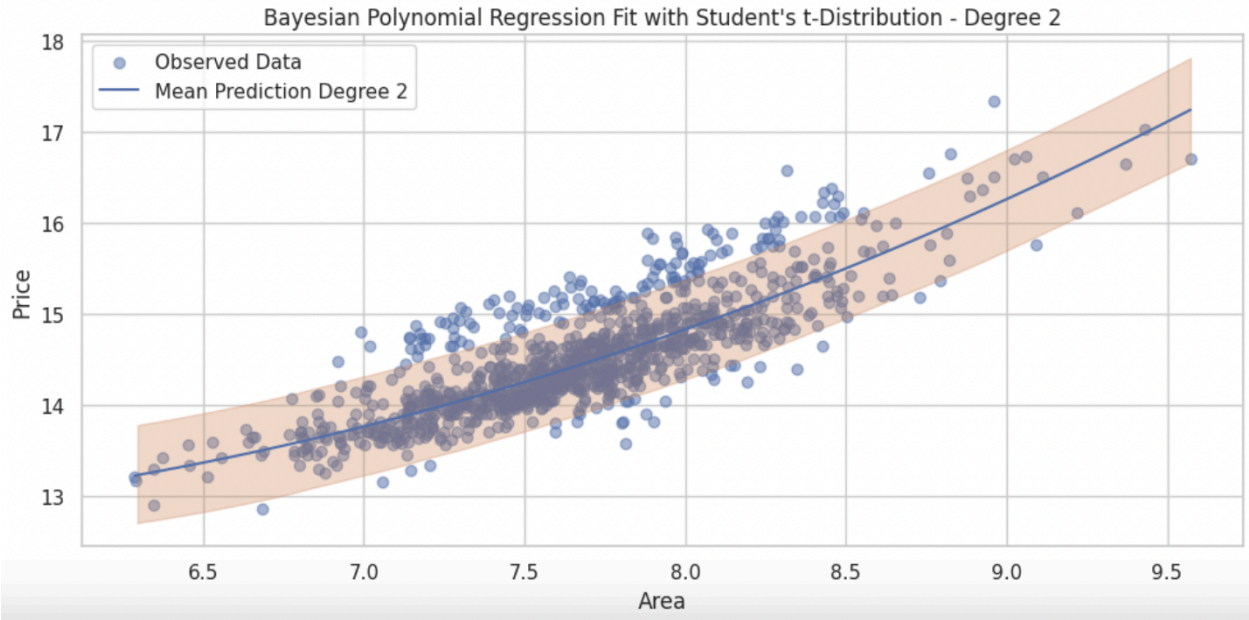
ν: A parameter that dictates the probability mass of the tails of the Student-T distribution. A smaller value means heavier tails, making the model more robust to outliers.

5. Evaluation and Comparison

Our sampling using Markov Chain methods appeared to show independent samples for degree 1 polynomials giving an accurate representation of underlying distribution, demonstrating that we can take our results to be reliable. However, degree 2 polynomials and higher had diagnostics

that pointed towards non-independent sampling, reflecting uncertainty regarding the underlying distribution.





We explored different Normal and t-distributed models of varying degree polynomials. Among these, the degree 2 polynomial model employing a Student's t-distribution emerged as the

standout performer, with the lowest LOO of 783, which estimates how well a model will perform on unseen data.

A model that performs well according to LOO metrics fits the existing data well and does so without becoming overly complex — a key to ensuring that it generalizes well to new data. The metrics of interest from the LOO analysis are the deviance LOO, which reflects the model's predictive accuracy, and the adequate number of parameters (p_loo), which gives an insight into the model's complexity.

This model's success can be attributed to the robustness of the Student's t-distribution in handling outliers and the flexibility of the degree 2 polynomial to capture the relationship between house area and price. However, this comes at the cost of poor sampling accuracy for higher degree polynomials, as observed, the safer option remaining using a simple degree 1 polynomial.

6. Conclusion

Our findings suggest that the best model for this data is with a quadratic equation using a Student T distribution to account for the high variance within the data. However, due to the poor diagnostic results regarding sampling, these results might not be the most accurate, and due to this uncertainty, as well as unnecessary model complexity of a higher degree polynomial, which might induce overfitting, the safer option and model of choice remains a simple degree 1 linear model. A predictive model can be built using the Most Likely Estimate value of our coefficients for the model. It might also be relevant to incorporate other factors to predict housing factors better while keeping a simple model.

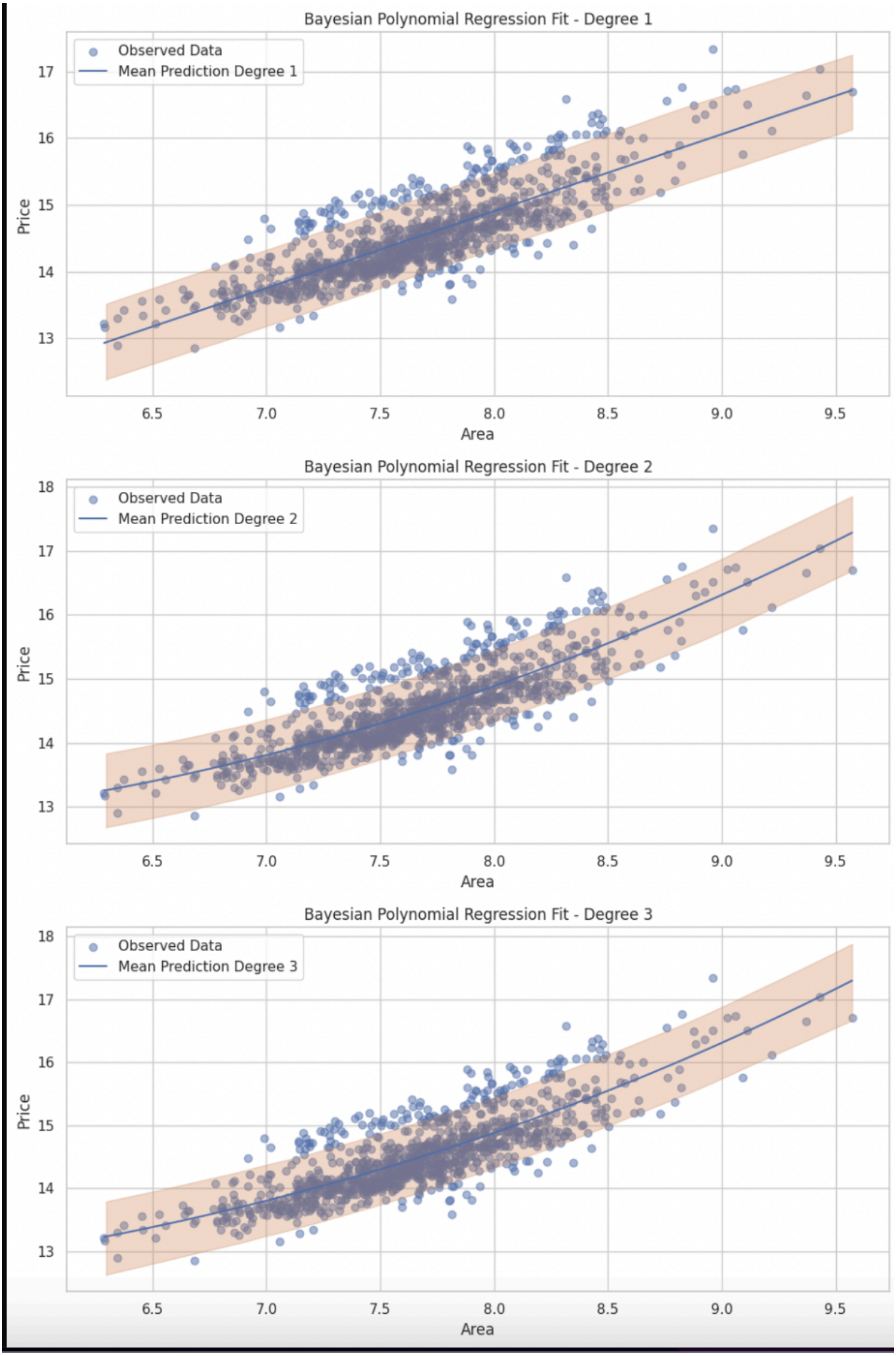
AI Statement: AI was used to help rewrite sentences to improve sentence structure and clarity.

References

Kulkarni, A. (2021). *Housing Prices in London*. Kaggle. Retrieved February 21, 2024, from https://www.kaggle.com/datasets/arnavkulkarni/housing-prices-in-london

Appendix

3rd Degree Polynomial fit attempt that could not be added to Google Colab because of technical issues.


Bayesian Polynomial Regression Fit - Degree 1
Observed Data
Mean Prediction Degree 1
Price
Area
Bayesian Polynomial Regression Fit - Degree 2
Observed Data
Mean Prediction Degree 2
Price
Area
Bayesian Polynomial Regression Fit - Degree 3
Observed Data
Mean Prediction Degree 3
Price
Area