Meeting Summary for the Live Professor Q&A on Regression at the 2025 Astrostatistics Summer School (06/03/2025)

1 Quick recap

The meeting focused on explaining logistic regression and its application as a classification tool, including discussions about model parameters, thresholds, and performance metrics. Ashley addressed various technical questions about model degeneracies, parameter estimation, and the handling of heteroscedastic data, while emphasizing the importance of model complexity selection and the role of physics-based insights in guiding model choice. The session concluded with discussions about error term independence, confidence interval calculations, and the ability of logistic regression to handle multiple classes, with some technical details deferred to follow-up communications on Slack.

2 Summary

2.1 Logistic Regression for Classification

Ashley explained the concept of logistic regression, focusing on how it transforms a linear model into a classification tool by using an activation function that outputs probabilities between 0 and 1. She clarified that the threshold for classifying data into categories (e.g., spiral vs. elliptical galaxies) is a hyperparameter that can be chosen based on the desired balance between precision and recall. Ashley also addressed Teodora's question about the limitations of using accuracy as a performance metric for logistic regression, explaining that accuracy is too discrete to effectively measure the model's performance in this context.

2.2 Dealing With Model Parameter Degeneracies

Ashley discussed the challenges of dealing with parameter degeneracies in models, explaining that in physics-based models, she prefers to capture full uncertainty rather than fix parameters, as it makes a physical statement. She noted that for data-driven models like neural networks, degeneracies are often accepted to achieve high accuracy, but for causal models, parameters might be fixed for better interpretation. Aleksandra raised a question about the applicability of these methods to heteroscedastic data, to which Ashley responded that the methods might still work but could require adjustments.

2.3 Modeling Uncertainty in Regression Analysis

Ashley discussed the flexibility of fitting methods like logistic and linear regression, noting that incorporating uncertainty into objective functions can handle heteroscedasticity and correlated noise. She emphasized that certain methods, such as correlation-based approaches, may not be reliable without preprocessing to address outliers. Ashley also addressed the challenge of distinguishing between instrumental systematics and true signals, suggesting techniques like null test cases and causal modeling to validate models. She concluded by explaining that error terms in linear models are not suitable for identifying outliers, as fitting uncertainties to each data point could be impractical and unbounded.

2.4 Model Complexity Selection Strategies

Ashley discussed the challenge of selecting an appropriate complexity for a model, particularly when fitting polynomial functions to data. She explained that the solution involves testing different polynomial orders on subsets of the data, noting that overly complex models tend to overfit and perform poorly on new data. Ashley also mentioned that while data-driven methods are commonly used, physics-based insights can sometimes guide the selection of an appropriate model complexity. The discussion concluded with a question about whether the independence of error terms affects least squares estimation, which Ashley clarified does not affect the estimation of parameters.

2.5 Statistical Estimation and Regression Techniques

Ashley discussed the independence of Epsilons and their impact on least squares estimates, noting that they do not affect these estimates but can influence hypothesis testing. She suggested consulting Slack for further clarification on correlated noise and data biases. Sagnik inquired about calculating confidence intervals for coefficient estimates in polynomial fits, and Ashley recommended a Bayesian approach, which would be covered in an upcoming lecture. Tom explained that bootstrapping could be a technique for calculating confidence intervals in linear models, and he emphasized the importance of distinguishing between linearity in parameters and inputs. Ashley also clarified that logistic regression can handle multiple classes beyond binary, using a cross-entropy score and a normalization technique to ensure outputs resemble probabilities. Sagnik asked about the nature of target points in local linear regression, and Ashley agreed to clarify

this on Slack due to potential misunderstanding.