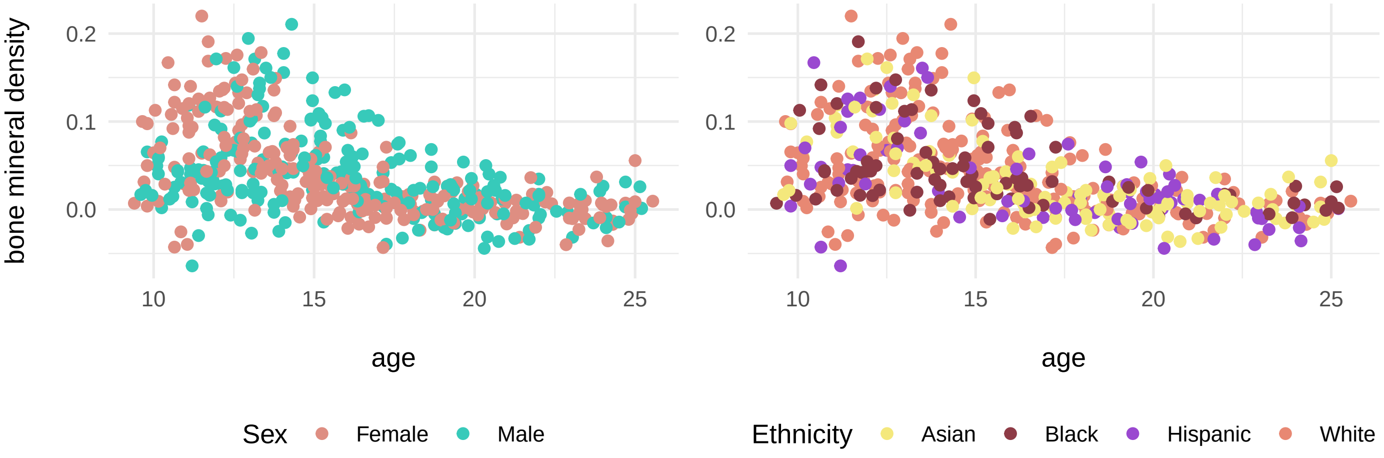
# Cross Validation for local polynomial linear regression

Alessandro Ferrera

The purpose of this report is to present the results of a Cross-Validation analysis conducted on a local polynomial linear regression model using the dataset `bone_mineral.csv`. This dataset contains information regarding bone mineral density, correlated with variables such as age, ethnicity, and sex.

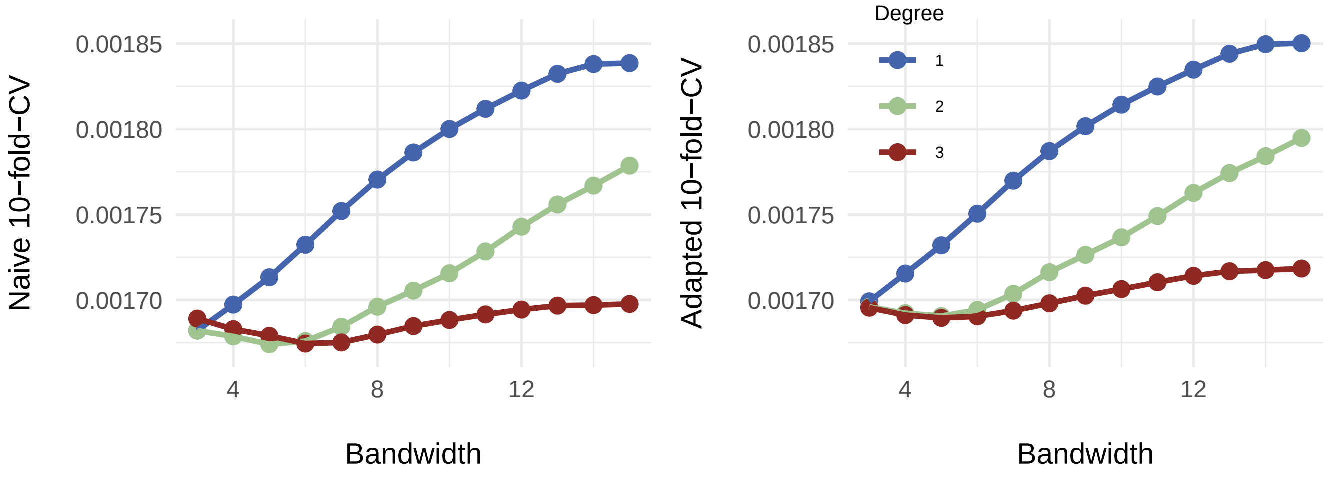
### Data Inspection

From a visual inspection, we can see that the bone mineral density is not correlated to the ethnicity or the sex. The data points are notably scattered throughout the plot, indicating no discernible pattern based on these factors. Consequently, it's not necessary to consider these variables in the model; thus, we will focus exclusively on age for the polynomial regression analysis.

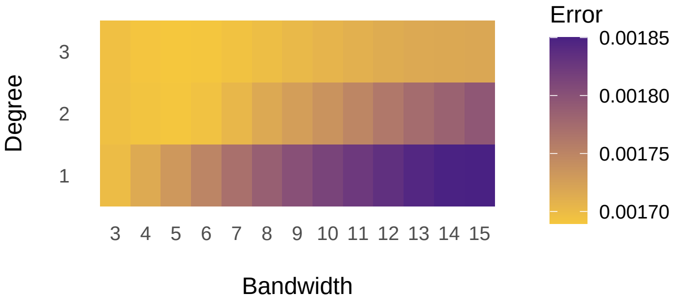


### Cross Validation

Consequently, we conducted a local polynomial regression analysis on the age, varying both the bandwidth and the polynomial degree. To identify the optimal model, we analyzed the K-fold Cross-Validation errors, with $K = 10$. Initially, this was executed using a "naive" approach, where the dataset was simply divided into training and testing sets. We then refined our methodology to ensure that both the training and testing folds adequately spanned the age range. All results presented in this report are consistent with this latter approach.

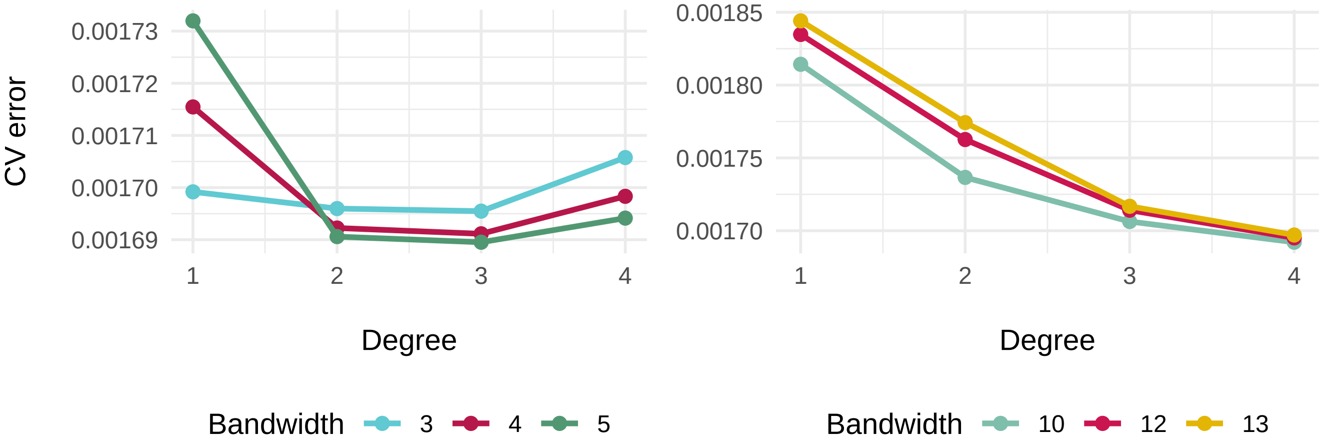


The plots show that the best model depends on the degree of polynomial $p$: the higher is $p$ the higher will be the optimal bandwidth $h$. Also this heat-map allows us to see this correlation. In the range we considered, the best model is the one with $p = 3$ and $h = 5$.



### Overfitting

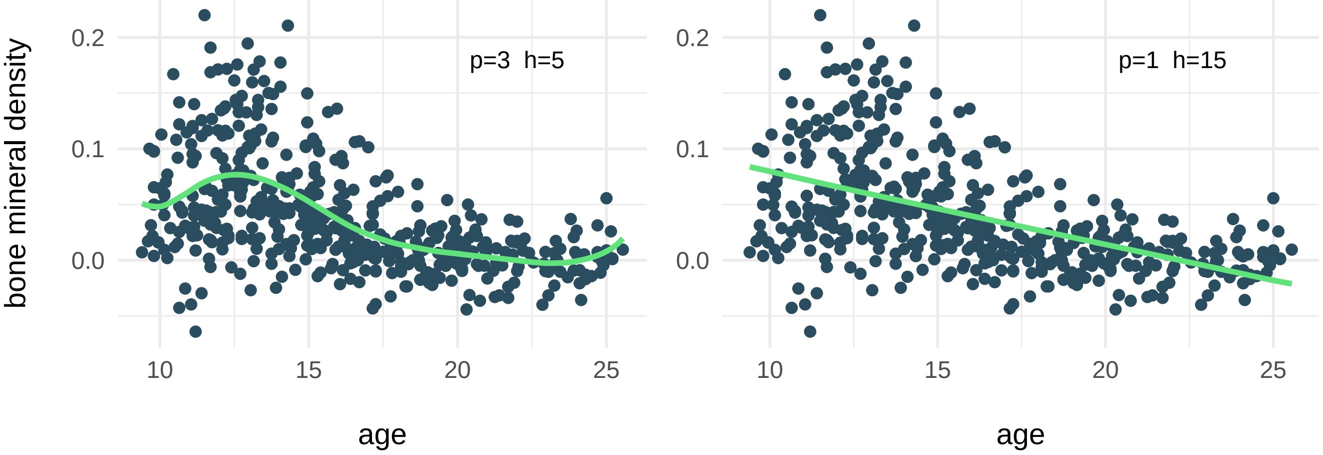
We also explored the issue of overfitting, a phenomenon characterized by the model's tendency to capture noise in the data rather than the underlying pattern.



This concern is illustrated in the accompanying plot, where a tighter bandwidth leads to a minimum in the cross-validation error at $p = 3$, after which the error begins to rise. In contrast, a wider bandwidth results in a continuously decreasing error. This suggests that with a larger bandwidth, the model becomes less flexible and thus requires a higher polynomial degree to adequately fit the data.

### Comparison between the best and worst model

We can finally show the difference between the best and the worst model we found.



### Conclusion

In conclusion, this analysis demonstrates the importance of selecting appropriate bandwidth and polynomial degree in local polynomial regression to avoid overfitting while still capturing the underlying trends in the data.