Guided Activity on Splines

Start Assignment

- Due Nov 5 by 11:59pm
- Points 100
- Submitting a file upload
- File Types pdf
- Available Oct 21 at 12am - Dec 20 at 11:59pm

Objective: This activity aims to explore non-linear relationships between a response variable and key predictors from a dataset of your choice using splines. You will select 2-3 predictors from the dataset, fit a linear model as a baseline, and then fit spline models to capture potential non-linear relationships. You will compare the performance of the models and interpret the results.

Suggested Datasets:

You may choose one of the following datasets for your analysis:

- **Wage Dataset** from the textbook *An Introduction to Statistical Learning*.
- **Auto Dataset** from the textbook *An Introduction to Statistical Learning*.
- **Bike Sharing Dataset** from the UCI Machine Learning Repository.
- **California Housing Data** from Kaggle.

Disclaimer: I have not explored these datasets extensively for this particular activity. If they are not interesting or the relationships between the response and predictors are not particularly non-linear, feel free to explore other datasets of your choice.

We will use the **California Housing Data** as an example to guide you through the process. The same steps apply to any of the datasets above.

Steps to Follow:

Step 1: Load and Understand the Dataset

Begin by downloading and preparing your dataset. Familiarize yourself with the response and predictor variables available in the dataset.

Example: California Housing Data

- **Response variable**: `MedHouseVal` (Median house value)
- **Predictors** include:
  - `MedInc`: Median income in block group.
  - `HouseAge`: Median house age in block group.

- `AveRooms`: Average number of rooms per household.
- `Population`: Block group population.
- `AveOccup`: Average number of household members.
- `Latitude` and `Longitude`: Coordinates of the block group.

Choose 2-3 predictors that you believe may have a non-linear relationship with the response variable. In this example, we will use `MedInc` (Median Income) and `HouseAge` (House Age).

Step 2: Data Exploration

Start by visualizing the relationship between your response variable and the chosen predictors. This helps to identify whether a linear model is suitable or if non-linear patterns exist.

Example:

- Plot the relationship between `MedHouseVal` (Median House Value) and two chosen variables:
  - `MedInc` (Median Income)
  - `HouseAge` (House Age)

This visualization will give you a sense of whether a linear model might be sufficient or if non-linearities are present.

Step 3: Fit a Linear Model (Baseline)

Before applying splines, fit a **linear regression model** to establish a baseline for comparison. This model will help you understand how the predictors influence the response under the assumption of a linear relationship.

- **Example**: Fit a linear model using `MedHouseVal` as the response and `MedInc` and `HouseAge` as the predictors.

Step 4: Fit Spline Models

Next, you will fit spline models to capture the potential non-linear relationship between the predictors and the response variable.

Step 4.1: Spline for Predictor 1 (e.g., Median Income)

- Use a **cubic spline** (or another type of spline of your choice) to model the relationship between the response variable and the first predictor.
- **Example**: Fit a spline model for `MedInc` to capture its non-linear relationship with `MedHouseVal`.

Step 4.2: Spline for Predictor 2 (e.g., House Age)

- Similarly, fit a spline for the second predictor to model its potential non-linear influence on the response.

- **Example**: Fit a spline model for `HouseAge` to capture its non-linear effect on `MedHouseVal`.

Feel free to experiment with different spline types or degrees for more flexibility. You may want to refer to the following papers for additional guidance on spline models:

- **Shape-Restricted Regression Splines with R Package splines2** (https://jds-online.org/journal/JDS/article/1243/info), by Wenjie Wang and Jun Yan.
- **A Review of Spline Function Procedures in R** (https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0666-3), by Aris Perperoglou, Willi Sauerbrei, Michal Abrahamowicz, & Matthias Schmid.

For more hands-on guidance on spline regression, the following resource is highly recommended: **Smoothing Spline Regression in R** (http://users.stat.umn.edu/~helwig/notes/smooth-spline-notes.html) by Nathaniel E. Helwig, University of Minnesota.

Step 5: Compare Linear and Spline Models

After fitting both the linear and spline models, compare their performance. Use metrics like **AIC** or **BIC** to determine which model fits the data better.

- **Example**: Compare the AIC/BIC values of the linear model and the spline models. This comparison will help determine whether introducing non-linear terms via splines improves model fit.

Step 6: Interpret the Results

1. **Interpretation of Coefficients**:
    - Analyze the coefficients from the spline models. How do the predictor variables (e.g., `MedInc` and `HouseAge`) affect the response variable in a non-linear fashion? Do the spline models reveal any interesting trends that the linear model missed?
2. **Model Comparison**:
    - How do the spline models improve over the linear model? Use the plots and AIC/BIC values to support your interpretation.
    - Look for differences in fit between the linear and spline models and reflect on whether the spline models better capture the complexity of the data.

Step 7: Conclusion

Summarize your findings:

- What did the spline model reveal about the relationships between your chosen predictors and the response variable?

- Is there clear evidence that splines provide a better fit compared to the linear model?
- Discuss any practical implications or insights from the non-linear model that may not have been obvious from a simple linear regression.

Deliverables:

You are required to present your results in a **Quarto** report. This report should include:

1. **Introduction**: A brief introduction to the dataset, your objective, and the chosen predictors.
2. **Data Exploration**: Visualizations and explanations of your exploratory analysis.
3. **Model Fitting**: Summarize the linear model and spline models you have fitted.
4. **Model Comparison**: Include AIC/BIC comparisons and performance metrics.
5. **Interpretation**: Discuss the interpretation of your spline models and how they differ from the linear models.
6. **Conclusion**: A final summary of your results, including the practical implications of your findings.

Ensure your Quarto report is well-structured, includes all relevant plots, and explains your analysis. Submit the report in PDF format.