6.6: Beyond Linearity

```
$ echo "Data Science Institute"
```

Introduction

In section 6.5 we saw how we can improve upon linear models by reducing their complexity and therefore the resulting variance of the estimates.

In this section we will look at extension of the linear model that do not require linearity.

- Polynomial Regression

- Step Functions

- Regression Splines

- Local Regression

- Generalized Additive Models

Polynomial Regression

The extension from linear regression to polynomial regression replaces the linear model

> $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with a polynomial model of degree d
>
> $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i$. The coefficient are easily
>
> \alert{estimated using least squares regression} with predictors $x_i, x_i^2, \ldots, x_i^d$.

- We do not often fit polynomials with degrees higher than 4 since they can be overly flexible and result in overfitting.

Exercises: Polynomial Regression

Open the Beyond Linearity Exercises Jupyter Notebook file.

- Go over the "Getting Started" section together as a class.

- Go over the "Polynomial Regression" section together as a class.

- 5 minutes for students to complete the questions.

- Questions should be completed at home if time does not allow.

Step Functions

Step functions involve breaking the range of the predictor X into sections with specified cut points $c_1, c_2, \ldots c_K$, called **knots**. Each section is then fit with a constant so our model becomes

$$
y_i = f(x_i) = \begin{cases} \beta_0 + \beta_1, & \text{if} \quad x_i \leq c_1 \\ \beta_0 + \beta_2, & \text{if} \quad c_1 \leq x_i \leq c_2 \\ \vdots \\ \beta_0 + \beta_K, & \text{if} \quad c_K \leq x_i \end{cases}
$$

We can use least squares to fit the coefficients.

Exercises: Step Functions

Open the Beyond Linearity Exercises Jupyter Notebook file.

- Go over the "Step Functions" section together as a class.

- 5 minutes for students to complete the questions.

- Questions should be completed at home if time does not allow.

Regression Splines

Piecewise polynomial regression is a combination of the previous two methods we have seen. We fit each of the K predictor sections with a polynomial of degree d.

$$y_i = f(x_i) = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \cdots + \beta_{d1}x_i^d, & \text{if} \quad x_i \leq c_1 \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \cdots + \beta_{d2}x_i^d, & \text{if} \quad c_1 \leq x_i \leq c_2 \\ \qquad\qquad\qquad \vdots \\ \beta_{0K} + \beta_{1K}x_i + \beta_{2K}x_i^2 + \cdots + \beta_{dK}x_i^d, & \text{if} \quad c_K \leq x_i \end{cases}$$

Again, the coefficients can be fit using least squares.

The **degrees of freedom** of a model is a measure of how flexible it is.

- The higher the degree of the polynomial and the more knots we have, the higher the degree of freedom.

- The higher the degrees of freedom the more flexible the model is.

Location and Number of Knot

Where to place knots?

- Regression splines are most stable in regions with a lot of knots.

- Place knots in places where function might vary the fastest.

- Although often knots are placed at uniform intervals for simplicity.

How many knots?

- Try out different numbers and see which looks best.

- Use cross-validation to find the value of K with the smallest RSS.

Constraints for Regression Splines

There are several useful constraints that can be used with regression splines to reduce their degrees of freedom, thereby reducing their flexibility and making the fit more stable.

- \alert{Continuity}: the fitted curve must be continuous.

- \alert{Continuous derivatives}: the first or second derivative must be continuous (this has the effect of requiring the curve to be smooth... no corners etc.)

- \alert{Natural spline}: the function must be linear in the region on the boundary (i.e. where X is smaller than the smallest knot or larger than the largest knot).

Exercises: Splines

Open the Beyond Linearity Exercises Jupyter Notebook file.

- Go over the "Splines" section together as a class.

Local Regression

Local regression involves fitting a function at a target point x_0 using only the nearby training observations.

1. Find the k training observations with x_i closest to x_0.

2. Assign weights K_{i0} to each of these points so that the farthest point has weight zero and the closest has the highest weight.

3. Fit a weighted least squares regression by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2.$$

4. The fitted function is then $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$

Choices for Local Regression

The most important choice is the \alert{span $s = k/n$ which is the fraction of points that are used to perform the local regression}.

- Controls the flexibility of the fit

    - small s results in a local, more wiggly fit

    - large s provides a global fit.

- Cross-validation can be used to choose s or it can be chosen directly.

There are other less important choices that can be made.

- How to define the weighting function K.

- What type of model to use such as constant, linear, or quadratic regression.

Exercises: Local Regression

Open the Beyond Linearity Exercises Jupyter Notebook file.

- Go over the "Local Regression" section together as a class.

- 10 minutes for students to complete the questions.

- Questions should be completed at home if time does not allow.

Generalized Additive Models

We have seen how we can extend simple linear regression beyond linearity. Now we will look at extensions for multiple linear regression. In the case of a $\alert{quantitative response}$ our linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

We can replace the linear $\beta_j x_{ij}$ with a non linear function $f_j\left(x_{ij}\right)$ to aquire the model

$$y_i = \beta_0 + f_1\left(x_{i1}\right) + f_2\left(x_{i2}\right) + \cdots + f_p\left(x_{ip}\right) + \epsilon_i$$

The functions $f_j\left(x_{ij}\right)$ can all have different forms including natural splines, polynomial regression, local regression etc. We can fit the model using least squares.

Generalized Additive Models

GAMs can also be used for qualitative Y. We will assume that Y is zero or one and $p(X) = \Pr(Y = 1|X)$. The logistic regression model is

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

We can incorporate non-linear relationships with functions $f_j(X_j)$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

Pros and Cons of GAMs

Pros

- We can model different non-linear relationships for each variable.

- Non-linear fits could provide more accurate predictions.

- The additive model allows for the examination of the effect of a single predictor while holding the rest constant.

Cons

- The interactions between variables are missed with the additive assumption. However, we can add additional predictors in the form of $X_i \times X_j$.

Exercises: GAMs

Open the Beyond Linearity Exercises Jupyter Notebook file.

- Go over the start of the "Generalized Linear Models" section together as a class.

- 10 minutes for students to complete the questions.

- Go over the rest of the "Generalized Linear Models" section together as a class.

- Questions should be completed at home if time does not allow.

References

Chapter 7 of the ISLR2 and ISLP books:

James, Gareth, et al. "Moving Beyond Linearity." An Introduction to Statistical Learning: with Applications in R, 2nd ed., Springer, 2021.

James, Gareth, et al. "Moving Beyond Linearity." An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.