Lab 7: Nonlinear Models

We will continue to use the `Wage` data set in the `ISLR` package to predict `wage` for $3,000$ mid-atlantic male workers.

```
library(ISLR)
library(tidyverse)
library(knitr)

str(Wage)
```

```
## 'data.frame':    3000 obs. of  11 variables:
##  $ year      : int  2006 2004 2003 2003 2005 2008 2009 2008 2006 2004 ...
##  $ age       : int  18 24 45 43 50 54 44 30 41 52 ...
##  $ maritl    : Factor w/ 5 levels "1. Never Married",..: 1 1 2 2 4 2 2 1 1 2 ..
##  $ race      : Factor w/ 4 levels "1. White","2. Black",..: 1 1 1 3 1 1 4 3 2 1
##  $ education : Factor w/ 5 levels "1. < HS Grad",..: 1 4 3 4 2 4 3 3 3 2 ...
##  $ region    : Factor w/ 9 levels "1. New England",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ jobclass  : Factor w/ 2 levels "1. Industrial",..: 1 2 1 2 2 2 1 2 2 2 ...
##  $ health    : Factor w/ 2 levels "1. <=Good","2. >=Very Good": 1 2 1 2 1 2 2 1
##  $ health_ins: Factor w/ 2 levels "1. Yes","2. No": 2 2 1 1 1 1 1 1 1 1 ...
##  $ logwage   : num  4.32 4.26 4.88 5.04 4.32 ...
##  $ wage      : num  75 70.5 131 154.7 75 ...
```

0.1 Polynomial Regression and Step Functions

1. Fit a degree-4 polynomial regression model predicting `wage` based on `age`. Inspect your model with the `summary` function. [**Hint**: you can use the `poly` function to create your polynomials in the model.]

2. One way we can choose the degree of our polynomial is through hypothesis testing. Fit polynomial models from linear to degree-5 of `wage` on `age`. We wish to choose the simplest model which is sufficient to explain the relationship between `wage` and `age`.

   To do this, we can use the `anova` function on our fitted models. This uses an F statistic to test the null hypothesis that a model $\mathcal{M}_1$ is sufficient to explain the data against a more complex model $\mathcal{M}_2$ (the alternative). Because our models are nested, we can compare all at once sequentially.

We will choose the simplest model that is still significantly different from the less complex model.

Use ANOVA (analysis of variance) to choose your polynomial regression model. Which model would you pick?

3. Choose your degree of polynomial using a cross validation approach. Do the chosen degrees match?

4. Fit a step function for `age` predicting `wage` with 4 cut points. You can use the function `cut` to change your quantitative variable into a categorical one. Let `cut` automatically choose the cut locations based on your data.

0.2 Regression Splines

To fit regression splines, we will use the `splines` library. The `bs` function generates a matrix of basis functions for regression splines (defaults cubic) based on a vector of knots or a specified degree of freedom. The `ns` function is the same for natural splines.

We can use either of these functions within the `lm` command:

```r
library(splines)
lm(y ~ bs(df = 5, degree = 2), data = df)
```

1. Fit `wage` on `age` using a cubic regression spline with knots at ages $25, 40, 60$.

2. Fit `wage` on `age` using a cubic regression spline with 6 degrees of freedom and knots chosen uniformly on the quantiles of the data (this is how `bs` does it by default).

3. Fit `wage` on `age` using a natural cubic regression spline with 6 degrees of freedom and knots chosen uniformly on the quantiles of the data.

4. Create a scatter plot of `wage` vs `age` with all three of your fitted splines overlayed as well as your chosen polynomial model (either by anova or CV). Comment on the shapes. [**Hint:** `predict` over a grid of `age` values might be helpful.]

0.3 GAMs

1. Fit a GAM using natural spline functions of `year` and `age`, treating `education` as a quantitative predictor. You can do this using either `lm` (least squares) or `gam` in the `gam` package (fit using back propagation).