Lab 6: Non-linear Modeling

First Name: ______________ **Last Name**: ______________ **NetID**: ______________

Lab 6 is due April 28 by 4:30pm in the homework box at 2nd floor of Rhodes Hall. For Lab 6, submit your code for the take home questions as well as your answers to the problems.

In this lab, we will learn how to use `R` to fit a **Generalized Additive Model** (GAM). We will consider both regression and classification on data set `Wage`. Our goal is on using ANOVA tests to select the right model with appropriate degrees of freedom.

The `Wage` data set consists of 12 variables (such as year, age, wage, and more) for 3000 people, and it is contained in library `ISLR`. Through out the lab we will treat `wage` as response variable and focusing on predictors `year, age, education`.

GAM for Regression Problems

Consider the task of fitting the model

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

Here `year` and `age` are quantitative variables, and `education` is a qualitative variable with five levels: `<HS, HS, <Coll, Coll, >Coll`, referring to the amount of high school or college education that an individual has completed. We fit the first two functions using various nonlinear functions, and the third function using a separate constant for each level via the usual dummy variable approach.

Splines

We can make use of the `splines` and `gam` libraries to fit various types splines in `R`. In particular:

- We use `bs()` from the `splines` library to generate the entire matrix of basis functions for splines with the specified set of knots. By default, cubic splines are produced.
- Similarly, we use `ns()` from the `splines` library for natural splines.
- For smoothing splines, we use `s()` from the `gam` library.

Here we fit a GAM to predict `wage` using natural spline functions of `year` and `age`, and specify that function of `year` should have 4 degrees of freedom, and that the function of `age` will have 5 degrees of freedom. Since this is just a big linear regression model using an appropriate choice of basis functions, we can simply do this using the `lm()` function. And the `lm()` function will automatically convert `education` into four dummy variables.

```
> library(ISLR)
> attach(Wage)

> library(splines)
> gam1=lm(wage~ns(year,4)+ns(age,5)+education,data=Wage)
```

Notice that without specification on `knots`, the default setting is to produce a spline with knots at uniform quantiles of the data.

The smoothing splines, unlike regression splines, cannot be expressed in terms of basis functions, and thus cannot be fitted directly using `lm()` via least squares. The `gam()` function from the `gam` library can be used to fit these more general types of splines.

For example, below we fit a GAM using smoothing splines with 4 degrees of freedom for `year`, and a smoothing spines with 5 degrees of freedom for `age`.

```
> library(gam)
> gam.m3=gam(wage~s(year,4)+s(age,5)+education,data=Wage)
```

Now let's see the fitted curves and their confidence intervals.

```
> par(mfrow=c(1,3))
> plot(gam.m3, se=TRUE,col="blue")
> plot.gam(gam1, se=TRUE, col="red")
```

Notice that the generic `plot()` function recognizes that `gam.m3` is an object of class `gam`, and invokes the appropriate `plot.gam()` method.

The two models, `gam.m3` and `gam1`, fitted using smoothing and natural splines, turn out to be quite similar. What do you observe from the above plots?

Holding age and education fixed, wage tends to increase slightly with year which might be the result of inflation. Holding education and year fixed, wage tends to be highest for intermediate values of age and lowest for the very young and very old. Holding year and age fixed, wage tends to increase with education that the more educated a person is, the higher their salary on average. All of these findings are intuitive.

Model selection using ANOVA

Now we have learned how to fit a specific nonlinear model. A more important question in practice is how to find the right nonlinear model with an appropriate amount of flexibility. Here we will explore how to do **model selection** using a technique called ANOVA.

In the above plots, the function of `year` looks rather linear. We can perform a series of ANOVA tests in order to determine which of these three models is the best:

($\mathcal{M}_1$) a GAM that excludes `year`,

($\mathcal{M}_2$) a GAM that uses a linear function of `year`,

($\mathcal{M}_3$) or a GAM that uses a smoothing spline function of `year`.

We use the `anova()` function, which performs an **analysis of variance** (ANOVA, using an F-test) in order to test the null hypothesis that a model $\mathcal{M}_1$ is sufficient to explain the data against the alternative hypothesis that a more complex model $\mathcal{M}_2$ is required. In order to use the `anova()` function, $\mathcal{M}_1$ and $\mathcal{M}_2$ must be *nested* models: the predictors in $\mathcal{M}_1$ must be a subset of the predictors in $\mathcal{M}_2$.

```
> gam.m1=gam(wage~s(age,5)+education,data=Wage)
> gam.m2=gam(wage~year+s(age,5)+education,data=Wage)
> anova(gam.m1, gam.m2, gam.m3, test="F")
```

What conclusion can you draw from the results of this ANOVA test? Therefore which of the 3 models do you think is preferred in this case? (Hint: check the p-values.)

The second model is preferred in this case since the p-value comparing the model 1 to model 2 is well less than 0.001, indicating model 1 is not sufficient to explain the data. The p-value of model 3 comparing to model 1 larger than 0.1 and therefore seems unnecessary. Hence, only model 2 appear to provide a reasonable fit to the data.

Lastly, we can take a look at the summary of the GAM fit.

```
> summary(gam.m3)
```

We can make predictions from `gam` objects, just like from `lm` objects, using the `predict()` function. Here we make predictions on the training set.

```
> preds=predict(gam.m2, newdata=Wage)
```

GAM for Classification Problems

Now we fit a GAM to the `Wage` data to predict the probability that an individual's income exceeds $250,000 per year. Our GAM has the form

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \times \texttt{year} + f_2(\texttt{age}) + f_3(\texttt{education})$$

where

$$p(X) = P(\texttt{wage} > 250 | \texttt{year, age, education})$$

In order to fit a logistic regression GAM, we use the `I()` function in constructing the binary response variable, and set `family=binomial`.

```
> gam.lr=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage)
> par(mfrow=c(1,3))
> plot(gam.lr,se=T,col="green")
```

Here the function for `year` is linear, f_2 is fitted using a smoothing spline with 5 degrees of freedom, and f_3 is fitted again using a different values for each dummy variables associated with `education`.

Now we can see that the last panel looks pretty weird, with very wide confidence intervals for level `<HS`. Check the the high earners in the `<HS` category using the command below and explain why.

```
table(education,I(wage>250))
```

No data point with wage > 250 has education level <HS.

Therefore, we exclude the category `<HS` and fit the model again.

```
> gam.lr=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage,
  subset=(education!="1. < HS Grad") )
```

Report your plots of the model.

Longer year of working lead to increase in probability of exceeding 250K wage which might be the result of inflation. The probability of exceeding 250K wage tends to be highest for intermediate values of age and lowest for the very young and very old, although the prediction is not very stable for the latter. For education level, the more educated a person is, the higher probability of exceeding 250K wage on average. All of these findings are intuitive.

For the plot please see attached page.

Take-Home Questions

1. Using other nonlinear models in a GAM

Write down the **R** command that fits a GAM of the form

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon,$$

where

- f_1 is a *polynomial* function of degree 6,
- f_2 is a *cubic spline* with 5 degrees of freedom (NOT a *natural* cubic spline),
- and f_3 is the same as before.

```
gam.m4=gam(wage~poly(year, 6)+bs(age,5)+education,data=Wage)
```

2. Model selection in GAMs for classification

We have used ANOVA to select models for the `year` variable. We can also do it for the `age` variable, and for a classification problem with **logistic regression**.

Consider the previous setting where we try to fit a logistic regression GAM of the form

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \times \texttt{year} + f_2(\texttt{age}) + f_3(\texttt{education})$$

Again we **use only the observations with more than a high school education**, and use ANOVA select from the following five models:

($\mathcal{M}_1$) a GAM that excludes `age` (i.e., $f_2 \equiv 0$)

($\mathcal{M}_2$) a GAM that uses a linear function f_2 of `age`

($\mathcal{M}_3$) a GAM that uses a smoothing spline f_2 of `age` with 2 degrees of freedom

($\mathcal{M}_4$) a GAM that uses a smoothing spline f_2 of `age` with 5 degrees of freedom

($\mathcal{M}_5$) a GAM that uses a smoothing spline f_2 of `age` with 8 degrees of freedom

Submit your code and report you ANOVA tests results. Based on the results, which model will you choose? Why?

I would choose either the second or the third model since their p-values comparing to model 1 are lower than 0.05, indicating that model 1 is not sufficient to explain the data. The p-values comparing the other models to model 1 are larger than 0.05 which seems unnecessary.

Therefore, although model 2 is less complexed than model 3,
either model 2 or model 3 appear to provide a reasonable fit to the data