SMM637 Answers - GAM

For this exercise consider the `Wage` data from `ISLR` library.

- Look at the help function of `Wage` for a description of the variables.
- Fit a GAM model with Gaussian distribution and identity link to predict `wage` using a TPRS functions of `age`. Also, include in the model `education` and `year` but without smooth functions.

```
library(ISLR)
library(mgcv)
data(Wage)
gam1 <- gam(wage ~ s(age) + education + year, data = Wage)
```

- Produce a summary of the gam fit and comment on the results.

```
summary(gam1)
```

  As for the parametric components, all coefficients are statistically different from zero. As for the interpretation, it is clear that having a higher education level, increases `wage`. In particular, on average the wage of a worker with an advanced degree is 63 unit greater than a wage of a worker with less than a higher school degree, keeping constant the reaming variables.

  As for the non-parametric component, we can see that the null hypothesis that the smooth function is zero is rejected. The edf is 4.8, suggesting a non linear effect of `age` on `wage`.

- Using the function `gam.check()`, produce a residual analysis and comment the findings.

```
gam.check(gam1)
```

  Normality assumption seem not to hold. Also, a problem of non-constant variance seems to be present.

- Plot the estimated smooth function of `age` with point-wise intervals. What do you notice?

```
plot(gam1, scale = 0)
```

  We can see that `wage` increases with `age`, reaching a maximum at 41/42 and then it starts decreasing. Also, the intervals seem to increase as age increases. So the smooth function is estimated with more uncertainty at higher values of `age`.

- Now fit a logistic regression GAM on the same dataset (this can be achieved by using a binomial distribution wtih logistic link). In order to do so we need to dichotomize our response variable `wage`. We can achieve this, for example, by setting `wage` to 1 if `wage`> 250 and zero otherwise. Look at the summary results, plot the estimated smooth unction and comment the results.

```
gam2 <- gam(I(wage >250) ~ year +s(age) + education, family = binomial, data = Wa
summary(gam2)
plot(gam2, scale = 0)
```

  None of the parametric components are significant. The smooth function for `age` is marginally significant. By dichotomizing the variable the information gets lost and the parameters are estimated with more uncertainty. The estimated smooth function shows to be non linear (inverted U shaped).