Homework: Non-Linear Regression

This homework sheet will test your knowledge of non-linear regressions using R.

a) Load the data named `College` from inside a the `ISLR` package. This dataset contains two columns denoting the percentage of faculty remembers with a Ph. D. degree, as well as the percentage of students finishing high school in the top 10 percent.

Plot a linear trend line between both variables. In comparison, also plot a LOESS smoothing in a separate plot. Do think a linear trend is an appropriate fit?

2

b) Repeat the above LOESS plots while trying different values for the smoothing intensity?

3

c) Use the same dataset from before. We now continue to calculate a trend line of our own as follows:

- Perform a polynomial regression of degree 3.
- Then, try to predict the percentage of students that finished high school in the top 10 percent while varying the percentage of faculty members with a Ph. D. between 0 % and 100 %.
- Visualize the line with predictions.
- Investigate finally which degree results into the best-fit model. Is there a model that is actually better than a linear trend?

4

d) Now that we introduced how to integrate quadratic terms into our model, we advance to adjust for interactions. For that purpose, we study how sales are linked to spending on TV and radio advertisements. Hence, estimate a linear model with least squares by including and excluding an iteration term. Afterwards, interpret the result.

Note: data set `Advertising` is available from `http://www-bcf.usc.edu/~gareth/ISL/data.html`

2

e) Load the dataset `Auto` from the `ISLR` library in order to study the relationship between weight and acceleration. When plotting both dimensions, we see a jump at a weight of around 3650, which is likely due to a different manufacturing type (limo vs. SUV).

As a result, we decide to perform a spline regression. Thus fit a quadratic polynomial to each region. In the end, visualize both curves nicely with a plot.

6

f) Use the same dataset as in the previous exercise. Create a polynomial regression of degree 3 that explains miles per gallon given an engine displacement. Visualize your trend line with 95 % confidence intervals.

Hint: the `predict(...)` routine has an additional parameter `se.fit` to calculate standard errors. These then just need to be scaled to the 95 % interval.

2

3

g) Now load the `Carseats` dataset from `ISLR` in order to visualize the different smoothings, namely, a linear trend, a generalized additive model and locally weighted scatterplot smoothing. Choose population on the x-axis and advertising spending on the y-axis.

Hint: you may prefer to deactivate standard error calculation in `geom_smooth(...)` because GAM does not support this feature.

2

h) Load the `Wage` dataset from `ISLR`. Determine the recommend number of degrees for a polynomial regression using ANOVA. Then plot your results.

Hint: `lapply(...)` allows to create multiple linear models at once. For convenience, read the documentation on `do.call(...)` when performing ANOVA for all models. In addition, the function `geom_smooth(...)` has an optional parameter `formula` which makes it easy to add custom smoothing functions.

3

i) Use once again the `College` dataset from `ISLR`. Use a GAM to predict whether a college is private or public based on the percentage of alumni that donate. Finally visualize your results. Also use an ANOVA test to show that smoothing splines are preferred over a linear effect.