# Rlab 6: Generalized Additive Models

STAT 410

**DUE: 12/02/2022 11:59pm**

## Homework Guidelines

***Please submit your answers on Canvas + Gradescope as a PDF with pages matched to question answers.***

Please prepare your solutions to this homework using R Markdown, which provides a way to include text, code, and figures in a single document. A template `.Rmd` file is available through Canvas. Make sure all solutions are clearly labeled, and utilize the pairing tool on Gradescope.

You are encouraged to work together, but your solutions, plots, and wording should always be your own. Come and see me during office hours or schedule an appointment when you get stuck and can't get unstuck.

## Relevant texts

HSAUR3 Ch. 10-10.3.1; Intro to Statistical Learning Ch. 7

## Data

The data for this homework assignment come from two sources: (1) the State of California **Energy Commission**, who maintain a record of the number of vehicles registered in the state by fuel type, and (2) the US Census, who provide, among other variables, both the population and number of people who work from home in every county. Henry used these two sources to create a file named `electric_WFH.csv` that has (i) the proportion of all vehicles that are electric, and (ii) the proportion of people who worked from home in 2018, each by county.

| Variable | Definition |
|---|---|
| County | County |
| prop_electric | Proportion of all vehicles that are electric |
| prop_WFH | Proportion of people who work from home |

The goal in this homework assignment will be to try to learn about any potential relationship between the proportion of vehicles that are electric in each county and the proportion of people who work from home.

### I. Linear regression [20 pts]

We will begin by analyzing the data using simple linear regression.

(1) [4 pts] **PEC** that fits a linear regression model for the proportion of vehicles that are electric as function of the proportion of people who work from home. Create a plot to assess the assumption of normally-distributed residuals. What do you notice?

(2) [4 pts] **PEC** that fits a different linear model, this time for the **log** of the proportion of vehicles that are electric as a function of the proportion of people who work from home. Again check the assumption of normally-distributed residuals. What do you notice now?

(3) [4 pts] Make a plot of the residuals from your model fit in (2) as a function of the fitted values (*hint: `plot(..., which = 1)`*). What does your plot suggest about the validity of the assumptions of **independent residuals** and/or a true **linear relationship** between the response and predictor?

(4) [4 pts] Make a plot of the log of the proportion of vehicles that are electric as a function of the proportion of people who work from home. Add a line that represents the fit from your model in (2) (*hint: you might find the function `abline()` useful*). Make a guess about why your plot in (3) looks the way it does.

(5) [4 pts] What is the proportion of people who work from home in San Diego County? Use your model from (2) to make a 95% confidence interval for the expected proportion of vehicles for counties with this proportion of people working from home. Make sure your prediction is for a proportion, *not a log-proportion*. How does your interval compare to the observed proportion of vehicles that are electric in San Diego County?

## II. Generalized Additive Model [20 pts]

(6) [4 pts] Use the `mgcv` package to fit a Generalized Additive Model (GAM) for the log of the proportion of vehicles that are electric as a flexible function of the proportion of people who work from home. Use the default smoothing basis of thin plate regression splines (i.e., `bs = "tp"`) and a basis dimension of 10. **PEC**. Make sure your coefficients match mine.

```
fit_gam$coefficients
```

```
##   (Intercept) s(prop_WFH).1 s(prop_WFH).2 s(prop_WFH).3 s(prop_WFH).4
##   -5.09007655   -0.39998424    0.01568659    0.01543831   -0.13729840
## s(prop_WFH).5 s(prop_WFH).6 s(prop_WFH).7 s(prop_WFH).8 s(prop_WFH).9
##   -0.05129185    0.21011465   -0.10667988    0.93605452    0.47079684
```

(7) [4 pts] Make a plot of residuals as a function of fitted values for your GAM in (6). How does it compare to the plot for the linear model in (3)? What does your new plot suggest about the validity of model assumptions?

(8) [4 pts] Make a new 95% CI for the expected proportion of vehicles that are electric in counties with the same proportion of people working from home as in San Diego County using your GAM from (6). How does your interval compare to the one you made in (5)? To the observed proportion of vehicles that are electric in San Diego County?

(9) [4 pts] Fit a new GAM for the log proportion of vehicles that are electric using cubic regression splines instead of thin plate splines and still uses a basis dimension of 10. **PEC**.

(10) [4 pts] Make a plot of *proportion of vehicles that are electric* as a function of the *proportion of people who work from home*. Add three curves representing the predicted proportion of vehicles that are electric according to the (1) linear, (2) GAM with thin plate regression splines and (3) GAM with cubic regression splines.