

Regression III: Lab 1

Dave Armstrong
University of Wisconsin – Milwaukee
Department of Political Science
e: armstrod@uwm.edu
w: www.quantoid.net/ICPSR.php

We discussed relatively simple solutions to estimate non-linear relationships between a continuous y and a categorical x . We also discussed local polynomial regression, spline regression models, smoothing splines and Generalized Additive Models. We'll work through a number of these concepts here.

1 Categorical Variables and Linearity

Do the following two commands in **R** to read in the data:

```
library(foreign)
dat <- read.dta("http://www.quantoid.net/lab1_nes.dta")
```

These are data from the American National Election Study. The variables included are:

pid	7-point party ID variable (Strong Dem=1, Strong Rep = 7)
pid3	3-point party ID variable (1=Dem, 2=Ind, 3=Rep)
demtherm	Democratic feeling thermometer
reptherm	Republican feeling thermometer
difftherm	Difference between democratic and republican thermometers
age	Survey respondent's age
income	23-category income variable
race	Un-recoded race variable
racerec	Recoded race variable
libcon	Liberal-conservative ideology (smaller = more liberal)

In all of the models you run below, make sure to control for **racerec** and **age** along with whatever variables I ask you to put in the model.

Let's try to assess the following hypotheses and present the evidence in favor or against them.

1. The 7-point party ID variable is linearly related to the democratic thermometer.
2. The 7-point party ID variable is not a significantly better predictor than the 3-point party ID variable.

How would you present these results?

2 Using Splines

Now, let's consider the effect of income on the democratic thermometer. This is categorical, in the sense that peoples' incomes are grouped into categories. You can make a numeric version of the dataset by typing

```
dat$newinc <- as.numeric(dat$income)
```

You will need to do this to use splines to investigate the relationship.

1. First, try to do this using dummy variables. Do we really need 22 parameters to characterize this relationship?
2. Using the `effect` command in the `effects` library, plot the effect of `income`.
3. Try to estimate the same relationship with a spline. How many knots should we have?
4. Use a generalized additive model to estimate the relationship. What does this tell you about the degrees of freedom used to estimate this relationship?
5. What sorts of things would you present if this were your work?

How would you interpret and present these results?

3 Interactions

Estimate a model of the difference in thermometer scores on `pid3`, `racerec`, and the interaction between `age` and `libcon` (also make sure you specify it so `R` includes the main effects).

1. Using the `effects` package, plot the fitted values of `difftherm` as a function of `age` holding `libcon` constant at different values and as a function of `libcon` holding `age` constant at different values.
2. Using the `DAMisc` package, plot the conditional coefficients and 95% confidence bounds for both variables involved in the interaction.
3. How do you interpret each of the plots?
4. What is the relationship between the plots in the first step and the plots in the second step?