Political Methodology III: Model Based Inference

Justin Grimmer
(Jens Hainmueller Based Slides)

May 9th, 2019

1. Learning the Regression Function

2. Generalized Additive Models
   - Setup
   - Fitting
   - Interpretation
   - Examples
   - Implementing GAMs in R

Linear Models

Recall the linear model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

Advantages:

- Simplicity
- Interpretability
- Easy to do inference

Downsides:

- Functional form assumptions (Linearity and Additivity)
- No learning

Linear Models

Recall the linear model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

Advantages:

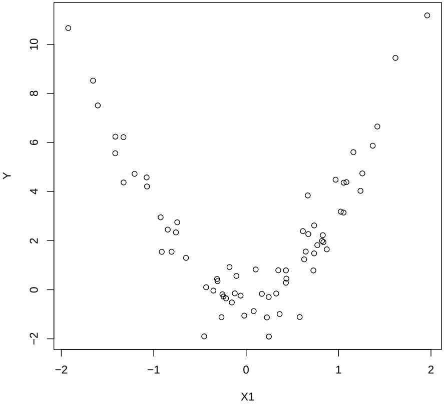
- Simplicity
- Interpretability
- Easy to do inference

Downsides:

- Functional form assumptions (Linearity and Additivity)
- No learning

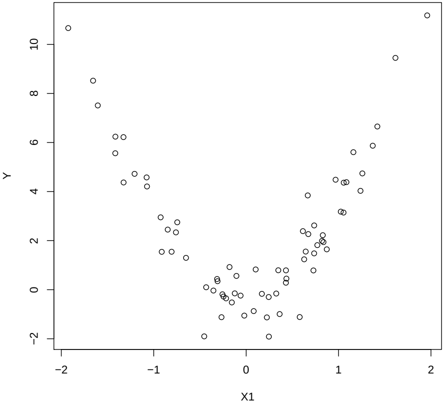
Nonlinearity

Linearity of the Conditional Expectation Function ($\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$) is a key assumption. Why?



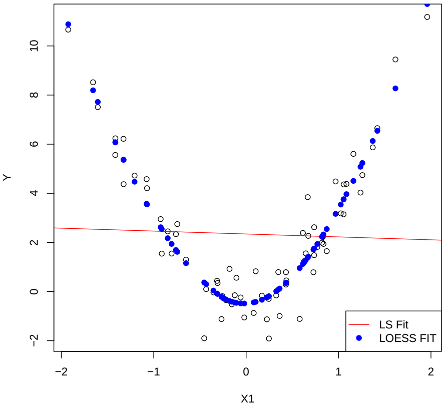
Nonlinearity

Linearity of the Conditional Expectation Function ($\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$) is a key assumption. Why?



Nonlinearity

Linearity of the Conditional Expectation Function ($\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$) is a key assumption. Why?



Nonlinearity

- If $E[Y|\mathbf{X}]$ is not linear in $\mathbf{X}$, $E[\mathbf{u}|\mathbf{X}] \neq 0$ for all values $\mathbf{X} = \mathbf{x}$ and $\hat{\beta}$ may be biased and inconsistent.
- Nonlinearities may be important but few social scientific theories offer any guidance as to functional form whatsoever.
  - Statements like "y increases with x" (monotonicity) are as specific as most social theories get.
  - Possible Exceptions: Returns to scale, constant elasticities, interactive effects, cyclical patterns in time series data, etc.
- Usually we employ "linearity by default" but we should try to make sure this is appropriate: detect non-linearities and model them accurately

Nonlinearity

- If $E[Y|\mathbf{X}]$ is not linear in $\mathbf{X}$, $E[\mathbf{u}|\mathbf{X}] \neq 0$ for all values $\mathbf{X} = \mathbf{x}$ and $\hat{\beta}$ may be biased and inconsistent.

- Nonlinearities may be important but few social scientific theories offer any guidance as to functional form whatsoever.
  - Statements like "y increases with x" (monotonicity) are as specific as most social theories get.
  - Possible Exceptions: Returns to scale, constant elasticities, interactive effects, cyclical patterns in time series data, etc.

- Usually we employ "linearity by default" but we should try to make sure this is appropriate: detect non-linearities and model them accurately

Nonlinearity

- If $E[Y|\mathbf{X}]$ is not linear in $\mathbf{X}$, $E[\mathbf{u}|\mathbf{X}] \neq 0$ for all values $\mathbf{X} = \mathbf{x}$ and $\hat{\beta}$ may be biased and inconsistent.

- Nonlinearities may be important but few social scientific theories offer any guidance as to functional form whatsoever.
  - Statements like "y increases with x" (monotonicity) are as specific as most social theories get.
  - Possible Exceptions: Returns to scale, constant elasticities, interactive effects, cyclical patterns in time series data, etc.

- Usually we employ "linearity by default" but we should try to make sure this is appropriate: detect non-linearities and model them accurately

Nonlinearity

- If $E[Y|\mathbf{X}]$ is not linear in $\mathbf{X}$, $E[\mathbf{u}|\mathbf{X}] \neq 0$ for all values $\mathbf{X} = \mathbf{x}$ and $\hat{\beta}$ may be biased and inconsistent.

- Nonlinearities may be important but few social scientific theories offer any guidance as to functional form whatsoever.
  - Statements like "y increases with x" (monotonicity) are as specific as most social theories get.
  - Possible Exceptions: Returns to scale, constant elasticities, interactive effects, cyclical patterns in time series data, etc.

- Usually we employ "linearity by default" but we should try to make sure this is appropriate: detect non-linearities and model them accurately

Diagnosing Nonlinearity

- For marginal relationships Y and X
  - Scatterplots with loess lines

- For partial relationships Y and X_1, controlling for X_2, X_3,...,X_k the regression surface is high-dimensional. We need other diagnostic tools such as:
  - Added variables plots and component residual plots

- Model diagnostics become rather difficult once we have many predictors

Diagnosing Nonlinearity

- For marginal relationships Y and X
  - Scatterplots with loess lines

- For partial relationships Y and X_1, controlling for X_2, X_3,...,X_k the regression surface is high-dimensional. We need other diagnostic tools such as:
  - Added variables plots and component residual plots

- Model diagnostics become rather difficult once we have many predictors

How should we deal with nonlinearity?

Given we have a linear regression model, our options are somewhat limited. We can partially address nonlinearity by:

- Breaking categorical or continuous variables into dummy variables (e.g. education levels)
- Including interactions
- Including polynomial terms
- Transformations such as logs

Alternative: Use a learning model that learns the functional form from the data (within a space of functions)

- Semi-parametric regression techniques: Generalized Additive Models (GAMs)
- Other machine learning methods: random forrest, boosted trees, ridge/lasso/elastic net regression, kernel based methods, etc.

Generalized Additive Models (GAM)

Take the linear model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

Ideally one would have a model that learns

$$y_i = f(x_{1i}, x_{2i}, x_{3i}) + u_i$$

but this is hard.

For GAMs, we maintain additivity, but instead of imposing linearity we allow flexible functional forms for each explanatory variable

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

where $s_1(\cdot), s_2(\cdot),$ and $s_3(\cdot)$ are smooth functions that are estimated from the data.

Generalized Additive Models (GAM)

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

- GAMs are semi-parametric, they strike a compromise between nonparametric methods and parametric regression
- GAMs are additive: create estimates of the regression surface by a sum of one-dimensional functions
- GAMs do not give you a set of regression parameters $\hat{\beta}$, but estimated functions $s_j(\cdot)$
- The functions can be summarized using a graphical summary of how $E[Y|X_,X_2,...,X_k]$ varies with X_1 (estimates of $s_j(\cdot)$ at every value of $X_{i,j}$)
- $s_j(\cdot)$ are usually estimated via back-fitting using locally weighted regression smoothers or various forms of splines (but many approaches are possible)
- Theory and estimation are somewhat involved, but they are easy to use

Generalized Additive Models (GAM)

- Assumption that the contribution of each covariate is additive is analogous to the assumption in linear regression that each component is estimated separately
- In OLS

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ji} + u_i$$

- GAMs assume similar additivity, but y_i is modeled as sum of arbitrary functions of the x-s

$$y_i = \beta_0 + \sum_{j=1}^{J} s_j(x_{ij}) + u_i$$

errors are assumed to be mean zero and constant variance

1. Learning the Regression Function

2. Generalized Additive Models
   - Setup
   - Fitting
   - Interpretation
   - Examples
   - Implementing GAMs in R

Generalized Additive Models (GAM)

- How do we learn the arbitrary functions $s_j(x_j)$?
- Consider the case where the Xs are independent, then we could simply estimate $s_j(x_j)$ by estimating a smooth of Y on each of the Xs separately
- Similarly in linear regression when the X-s are uncorrelated, the partial regression slopes are identical to the marginal regression slopes
- Since the Xs are usually correlated, we need to remove the effects of other predictors
- Solution: Backfitting algorithm that finds each function while controlling for the effects of the other Xs

Backfitting

- Suppose that we have a two predictor additive model:

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + u_i$$

- If we knew the partial regression function s_2 but not s_1 we could re-arrange

$$y_i - s_2(x_{2i}) = \beta_0 + s_1(x_{1i}) + +u_i$$

and use this to estimate s_1 by smoothing $y_i - s_2(x_{2i})$ against x_1.

- Knowing one partial function allows us to find the other
- We don't know s_2, but we can approximate it with an initial guess and then proceed iteratively → backfitting!

Backfitting

1. Express variables in mean-deviation form to eliminate intercepts (partial regressions sum to zero)

2. Run simple linear regression to get initial estimates of each partial regression function

$$y_i - \bar{y} = \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \epsilon_i$$

$$y_i^\star = \beta_1 x_{1i}^\star + \beta_2 x_{2i}^\star + \epsilon_i$$

3. Use estimates as step 0 in iterative estimation procedure

$$\hat{s_1}^{(0)} = \hat{\beta}_1 x_{1i}^\star \text{ and } \hat{s_2}^{(0)} = \hat{\beta}_2 x_{2i}^\star$$

4. Then we find partial residuals for x_1 which removes y from its linear relationship with x_2, but retains relationship between y and x_1. Partial residuals for x_1 are

$$\epsilon_{i[1]}^{(1)} = y_i^\star - \hat{\beta}_2 x_{2i}^\star = \epsilon_i + \hat{\beta}_1 x_{1i}^\star$$

and we do the same for x_2

Backfitting

5. Now we can smooth the partial residuals against their respective Xs to provide an updated estimate of $s()$.

$$\hat{s_k}^{(1)} = \text{smooth}[\epsilon_{i[k]}^{(1)} \text{ on } x_{ik}]$$

   where any reasonable smoother could be used (e.g. a loess or spline)

6. Iterate in finding updated estimates of the functions until the partial functions converge, ie. they no longer change from one iteration to the next

Key result: We now have estimates of $s_k(x_{ij})$ for every value of x_j and we have reduced the multiple regression problem into a series of two-dimensional partial regression functions.

This makes interpretation easy: we can plot functions which show the partial effects of each x on y (controlling for other Xs)

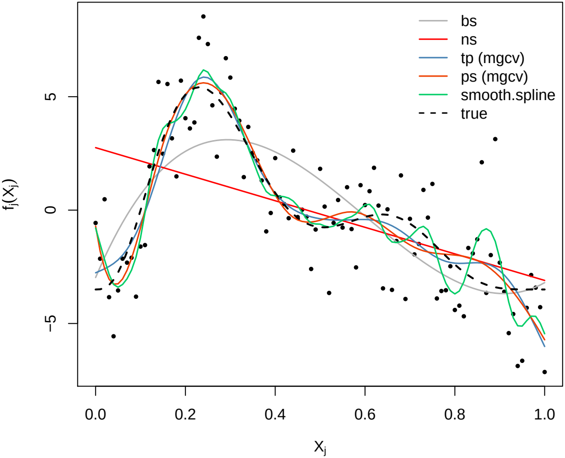
1. Learning the Regression Function

2. Generalized Additive Models
   - Setup
   - Fitting
   - Interpretation
   - Examples
   - Implementing GAMs in R

Interpretation: GAM Partial Regression Plots

- A plot of of x_j versus $s_j(x_j)$ shows the relationship between x_j and y holding constant the other variables in the model
- Since y is in mean deviation form, the smooth term $s_j(x_j)$ is also centered and thus each plot represents how y changes relative to its mean with changes in x
- Interpretation of scale of GAM plots is also straightforward
  - Value of 0 on Y-axis is the mean of y
  - Deviations from zero are deviations from the mean and we can add the mean to obtain the fitted values

Which Smoother to Use?



Need to find the right balance between under and over-smoothing

Generalized Additive Models (GAM)

The GAM approach can be extended to allow interactions ($s_{12}(\cdot)$) between explanatory variables, but this eats up degrees of freedom so you need a lot of data (often thin plate splines are used here).

$$y_i = \beta_0 + s_{12}(x_{1i}, x_{2i}) + s_3(x_{3i}) + u_i$$

It can also be used for hybrid models where we model some variables as parametrically and other with a flexible function:

$$y_i = \beta_0 + \beta_1 x_{1i} + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

Generalized Additive Models (GAM)

The GAM approach can be extended to allow interactions ($s_{12}(\cdot)$) between explanatory variables, but this eats up degrees of freedom so you need a lot of data (often thin plate splines are used here).

$$y_i = \beta_0 + s_{12}(x_{1i}, x_{2i}) + s_3(x_{3i}) + u_i$$

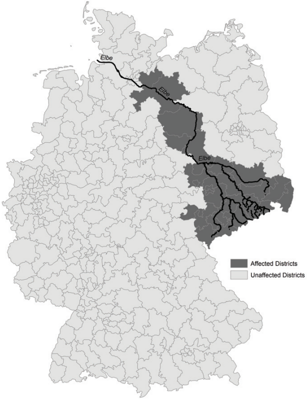
It can also be used for hybrid models where we model some variables as parametrically and other with a flexible function:

$$y_i = \beta_0 + \beta_1 x_{1i} + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

1. Learning the Regression Function

2. Generalized Additive Models
    - Setup
    - Fitting
    - Interpretation
    - Examples
    - Implementing GAMs in R

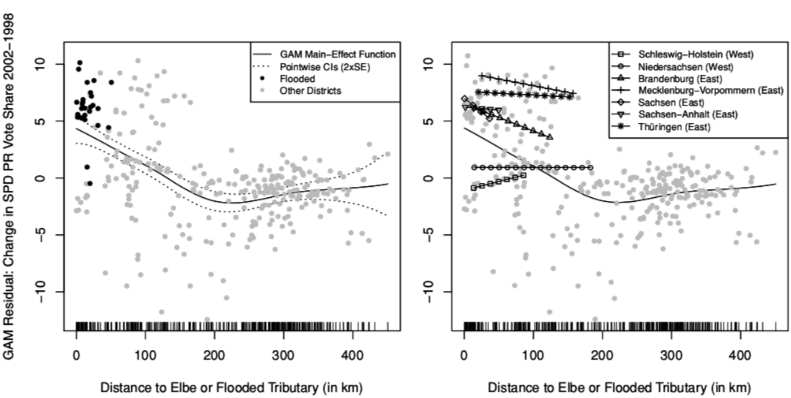
Effect of Disaster Aid on Vote Share



Bechtel and Hainmueller. 2011. "How Lasting is Voter Gratitude? An Analysis of the Short- and Long-term Electoral Returns to Beneficial Policy". American Journal of Political Science.

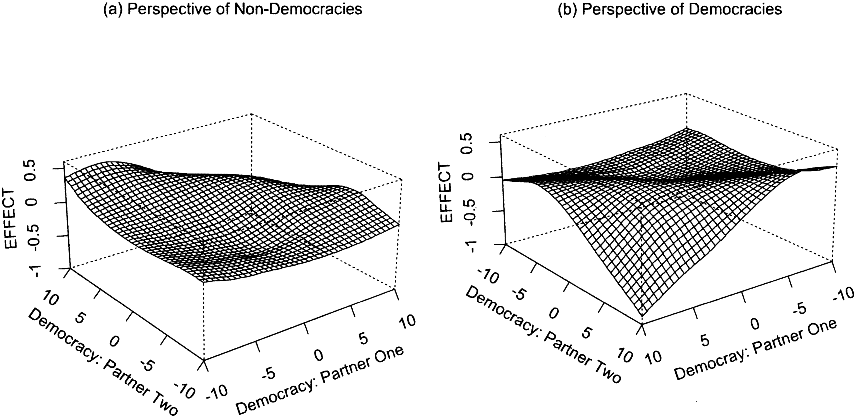
Effect of Disaster Aid on Vote Share

FIGURE 5 **Change in SPD PR Vote Share 2002–1998 and Distance to Elbe**



Note: The changes in SPD PR vote share from the 1998 to 2002 elections are modeled with a General Additive Model (GAM) that includes the full set of covariates and a (back-fitted) smoothing spline for the distance to the Elbe River or the closest flooded tributary. The left figure plots the partial deviance residuals from the GAM fit against the distance to the Elbe with the main-effect function superimposed (and twice standard error confidence envelopes). Directly flood-affected districts (*Flooded* = 1) are highlighted in black. The right figure shows the same plot with superimposed local linear regression lines that visualize the conditional relationship between SPD vote gains and the distance to the Elbe within each of the land regions that have at least one directly affected district.

GAM Fit to Diadic Democracy and Militarized Disputes



1 Learning the Regression Function

2 Generalized Additive Models
- Setup
- Fitting
- Interpretation
- Examples
- Implementing GAMs in R

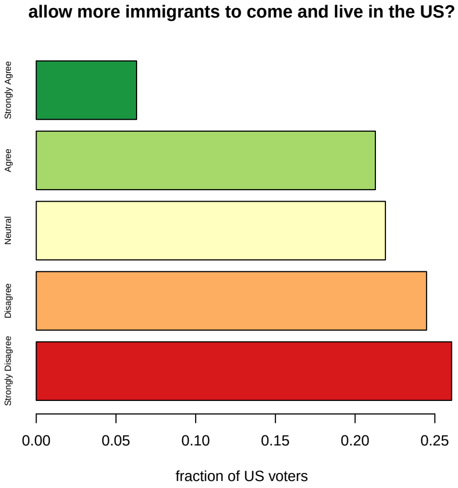
Generalized Additive Models (GAM)

- Workhorse function is *gam()* in *mgcv* package
- The formula takes the same form as the *glm* function except now we have the option of having parametric terms and smoothed estimates
- Smooths will be fit to any variable specified with the *s(x)* argument (can also use other smoothers *te, ti or t2*)
- Example: *gam.out <- gam(y∼x1+s(x2)+s(x3),data=data)* (add *family* to use with any standard GLM)
- Can use *plot(gam.out)*, *predict(gam.out)*, and *summary(gam.out)* methods to interpret results
- The summary function returns tests for each smooth, the degrees of freedom for each smooth, and an adjusted R- square for the model.
- Statistical inference and hypothesis testing is based on the deviance with approximate empirical degrees of freedom

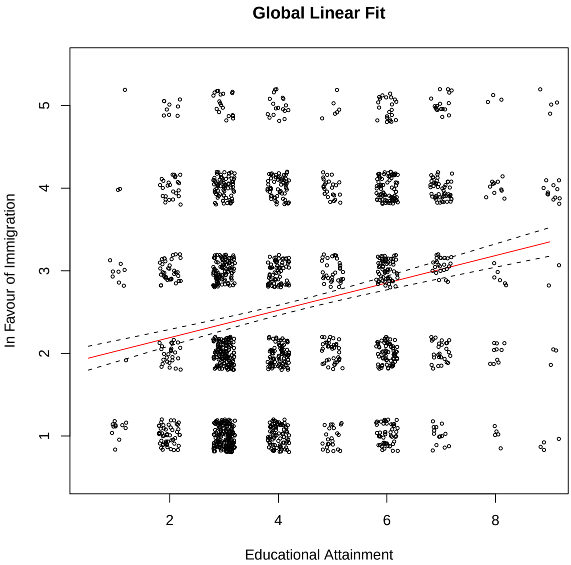
Example: Attitudes Towards Immigration

- Outcome: Pro-Immigration Attitudes
    - 1 Strongly opposed to increase in immigration
    - .
    - 5 Strongly in favour of increase in immigration
- Highest Educational Attainment:
    - 1 Less than high school
    - 2 Some high school, no diploma
    - 3 Graduated from high school- Diploma or equivalent (GED)
    - 4 Some college, no degree
    - 5 Associate degree (AA, AS)
    - 6 Bachelor's degree
    - 7 Master's degree
    - 8 Professional degree (MD, DDS, LLB, JD)
    - 9 Doctorate degree
- Age of Respondent
    - 18 years to 93 years

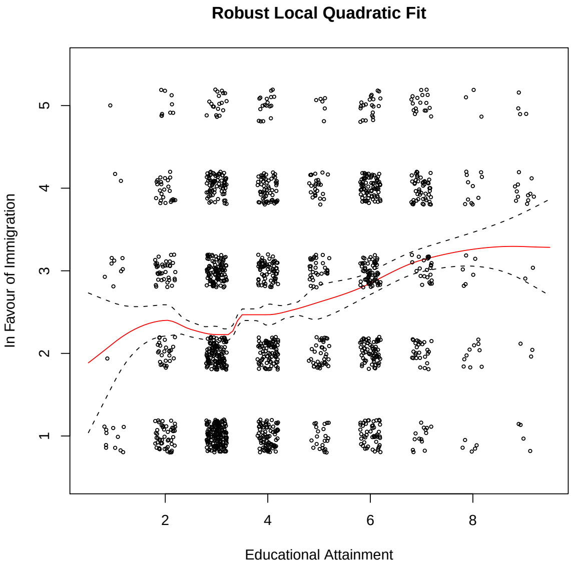
Attitudes Towards Immigration (2008)



Immigration Attitudes and Education



Immigration Attitudes and Education



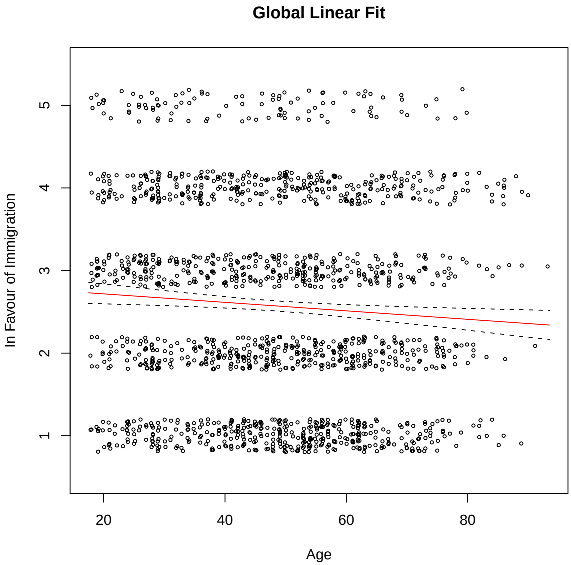
Immigration Attitudes and Education

```
[fontsize=\footnotesize, frame=single, label=R Code]
> mod0 <- lm(imgpro5mod1˜ppeduc, data=d)
> coeftest(mod0)

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.859180   0.081894 22.7022 < 2.2e-16 ***
ppeduc      0.165658   0.017636  9.3929 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

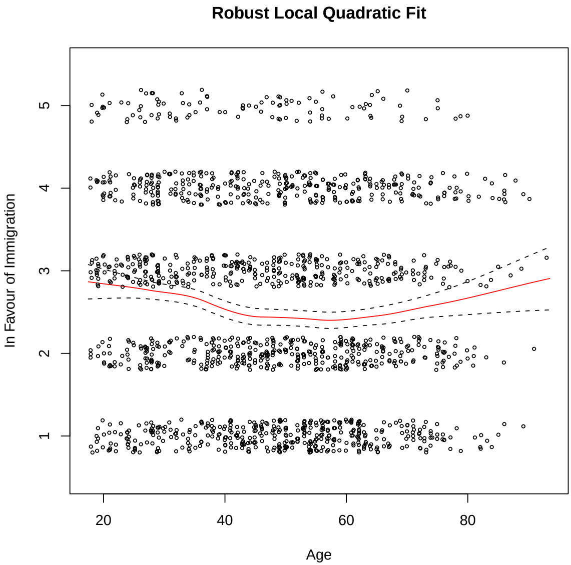
Immigration Attitudes and Education

```
[fontsize=\footnotesize, frame=single, label=R Code]
> mod0a <- lm(imgpro5mod1~factor(ppeduc), data=d)
> coeftest(mod0a)

t test of coefficients:

                Estimate Std. Error t value  Pr(>|t|)    
(Intercept)     2.173913   0.253720  8.5682 < 2.2e-16 ***
factor(ppeduc)2 0.261461   0.272848  0.9583 0.3380747    
factor(ppeduc)3 0.096465   0.259456  0.3718 0.7100945    
factor(ppeduc)4 0.324500   0.262820  1.2347 0.2171310    
factor(ppeduc)5 0.461008   0.275907  1.6709 0.0949432 .  
factor(ppeduc)6 0.673605   0.263864  2.5528 0.0107779 *  
factor(ppeduc)7 1.059170   0.274784  3.8546 0.0001206 ***
factor(ppeduc)8 0.740373   0.326614  2.2668 0.0235363 *  
factor(ppeduc)9 1.186087   0.351565  3.3737 0.0007595 ***
```

Immigration Attitudes and Age



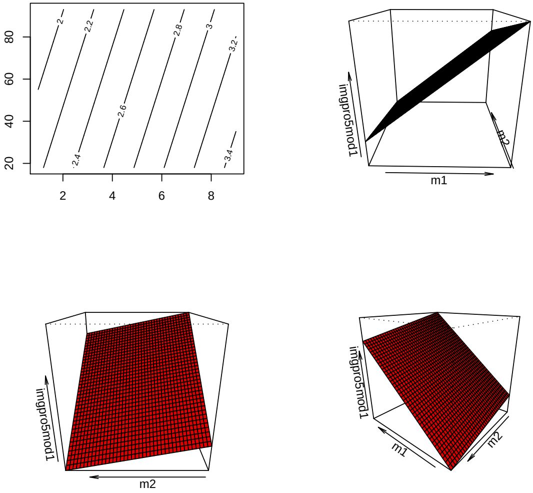
Immigration Attitudes and Age



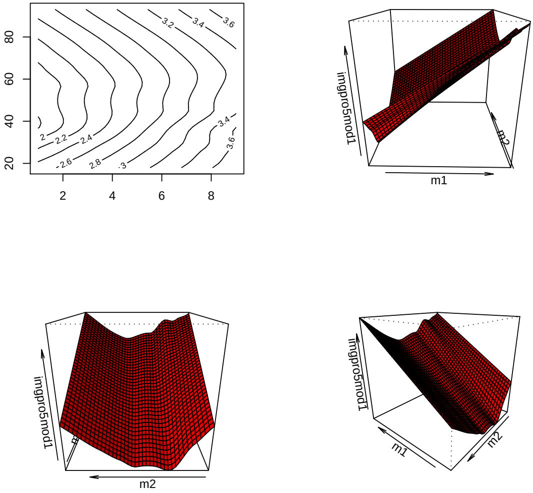
Immigration Attitudes and Age

```
[fontsize=\footnotesize, frame=single, label=R Code]
> mod1 <- lm(imgpro5mod1~ppage, data=d)
> coeftest(mod1)
t test of coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  2.8201106  0.0957797  29.444 < 2.2e-16 ***
ppage       -0.0051324  0.0018772  -2.734  0.006326 **
>
> mod1a <- lm(imgpro5mod1~ppage+I(ppage^2), data=d)
> coeftest(mod1a)
t test of coefficients:
               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  3.60493245  0.23929787 15.0646 < 2.2e-16 ***
ppage       -0.04123139  0.01026562 -4.0165 6.184e-05 ***
I(ppage^2)   0.00036702  0.00010262  3.5763 0.0003589 ***
```

Immigration Attitudes, Education and Age



Immigration Attitudes, Education and Age



Immigration Attitudes, Education and Age

Go through GAM Code

Further Readings on More Flexible Regression Methods

- Beck, N. and Jackman, S. 1998. Beyond Linearity by Default: Generalized Additive Models. *American Journal of Political Science*.
- Wood (2003). "Thin plate regression splines." *Journal of the Royal Statistical Society: Series B*.
- Hastie, T.J. and Tibshirani, R.J. 1990. *General Additive Models*.