# Generalized Linear Models

Seminar in Criminology, Research and Analysis– Crim 7301
Week 4, 9/15/16
Andrew Wheeler

# Class Overview

- Logistic and Poisson regression
- Graphing Effects
- Non-linear effects in models

# Generalized Linear Models

- Can be written as

$$\mathbb{E}[Y] = f(X_1, X_2, \dots.)$$

Where $f$ is an anonymous function

# Generalized Linear Models

- Motivated when the *dependent* variable has a particular distribution

  - Poisson – data are integer counts, e.g. 0, 1, 2, … and do not have an upper bound. (So Likert data that is 1 to 5 would not be appropriate). Zero values should be a possibility.

  - Logistic – data are 0-1 (binary) (other variants include more than one outcome category (multinomial), or ordered levels (like Likert data)

Distribution of the *independent* variables does not matter!

# Generalized Linear Models

- Poisson regression, the function is the *exponential*, so can write as:

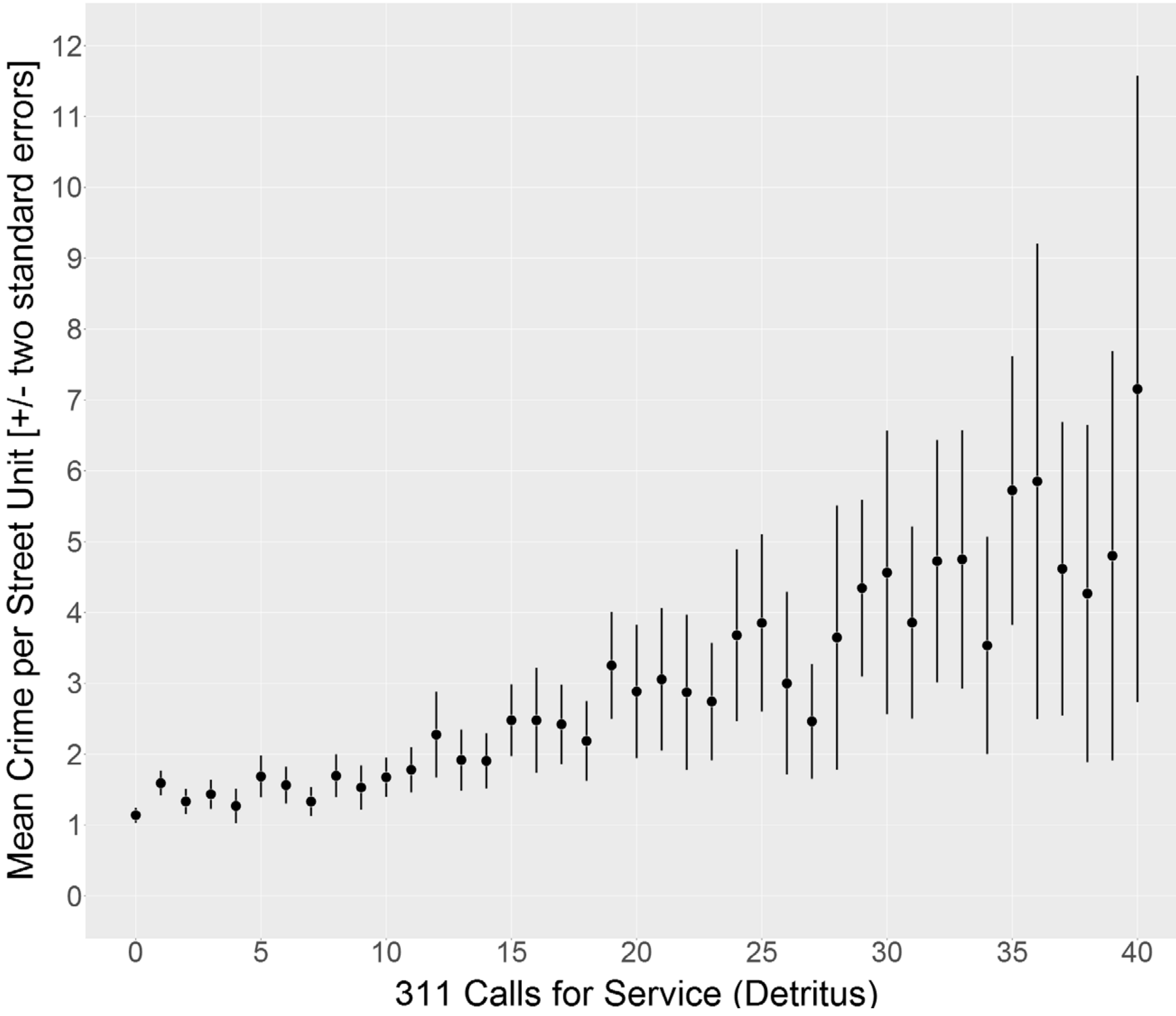
$$\mathbb{E}[Y] = \exp(\beta_0 + \beta_1 X)$$

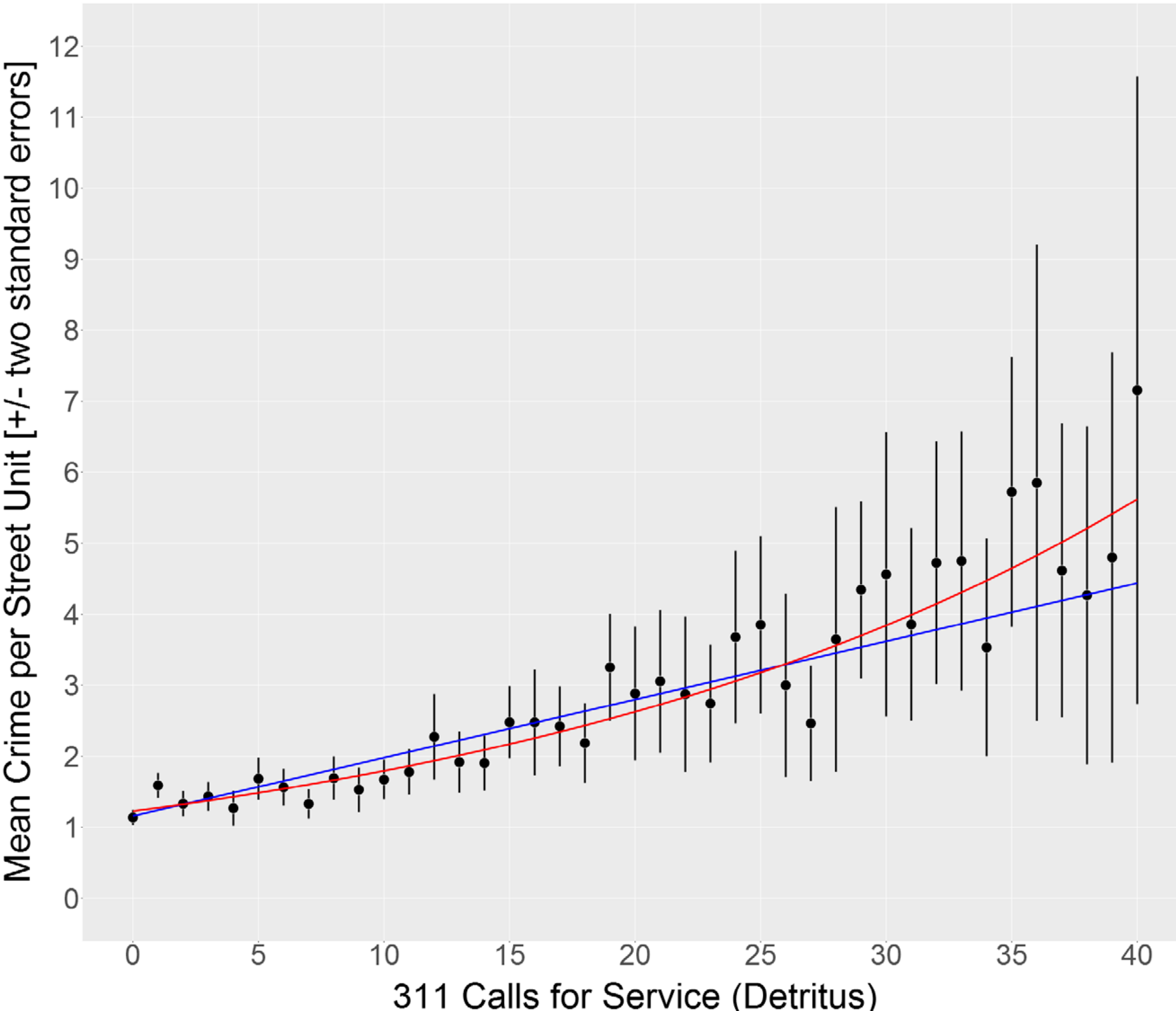
- Or equivalently

$$\log(\mathbb{E}[Y]) = \beta_0 + \beta_1 X$$

# Generalized Linear Models

- But, also forces the functional form of the relationship to be whatever the link function is, e.g. exponential in Poisson regression


Mean Crime per Street Unit [+/- two standard errors]
311 Calls for Service (Detritus)
0
1
2
3
4
5
6
7
8
9
10
11
12
0
5
10
15
20
25
30
35
40



Mean Crime per Street Unit [+/- two standard errors]
0
1
2
3
4
5
6
7
8
9
10
11
12
0
5
10
15
20
25
30
35
40
311 Calls for Service (Detritus)

# Generalized Linear Models

- Logistic regression the function is more complicated:

$$\mathbb{E}[Y] = \text{logistic}(\beta_0 + \beta_1 X)$$

- Where

$$\text{logistic}(x) = \frac{1}{1 + e^{-x}}$$

- To simplify things, just define the inverse logistic function (ie the logit) and can write as:

$$\text{logit} = \ln[\frac{\text{logistic}(x)}{1 - \text{logistic}(x)}]$$

$$\text{logit}(\mathbb{E}[Y]) = \beta_0 + \beta_1 X$$

Or

$$\text{logisitic}^{-1}(\mathbb{E}[Y]) = \beta_0 + \beta_1 X$$

# Graphing Effects

- To compare linear and non-linear models, just pick a particular set of inputs, and see what the different models predict

- Example, with 0/1 data, the Linear Probability Model

$$\mathbb{E}[\text{Recidivism}] = 0.4 + 0.4(X_1) - \ 0.2(X_2)$$

- Vs Logistic Regression

$$\mathbb{E}[\text{Recidivism}] = \text{logistic}[-0.4 + 1.8(X_1) - 0.8(X_2)]$$

# Graphing Effects

- When x1 = 1 in linear model and x2 = 0

$$\mathbb{E}[\text{Recidivism}] = 0.4 + 0.4(1) - 0.2(0) = 0.6$$

- In the logistic model

$$\mathbb{E}[\text{Recidivism}] = \text{logistic}[-0.4 + 1.8(1) - 0.8(0)] = 0.65$$

# Graphing Effects

- When to choose linear over generalized linear?
  - It is defacto standard to choose logistic for 0/1 data and Poisson for count data
  - Linear is not a bad substitute *if* the predictions are mostly within permissible ranges.
  - Some types of models need to be linear (some structural equation models, ARIMA with endogenous lags, certain panel data models)

# Non-linear effects

- Can have non-linear effects in OLS
- Frequently modelled as *polynomials*, e.g.

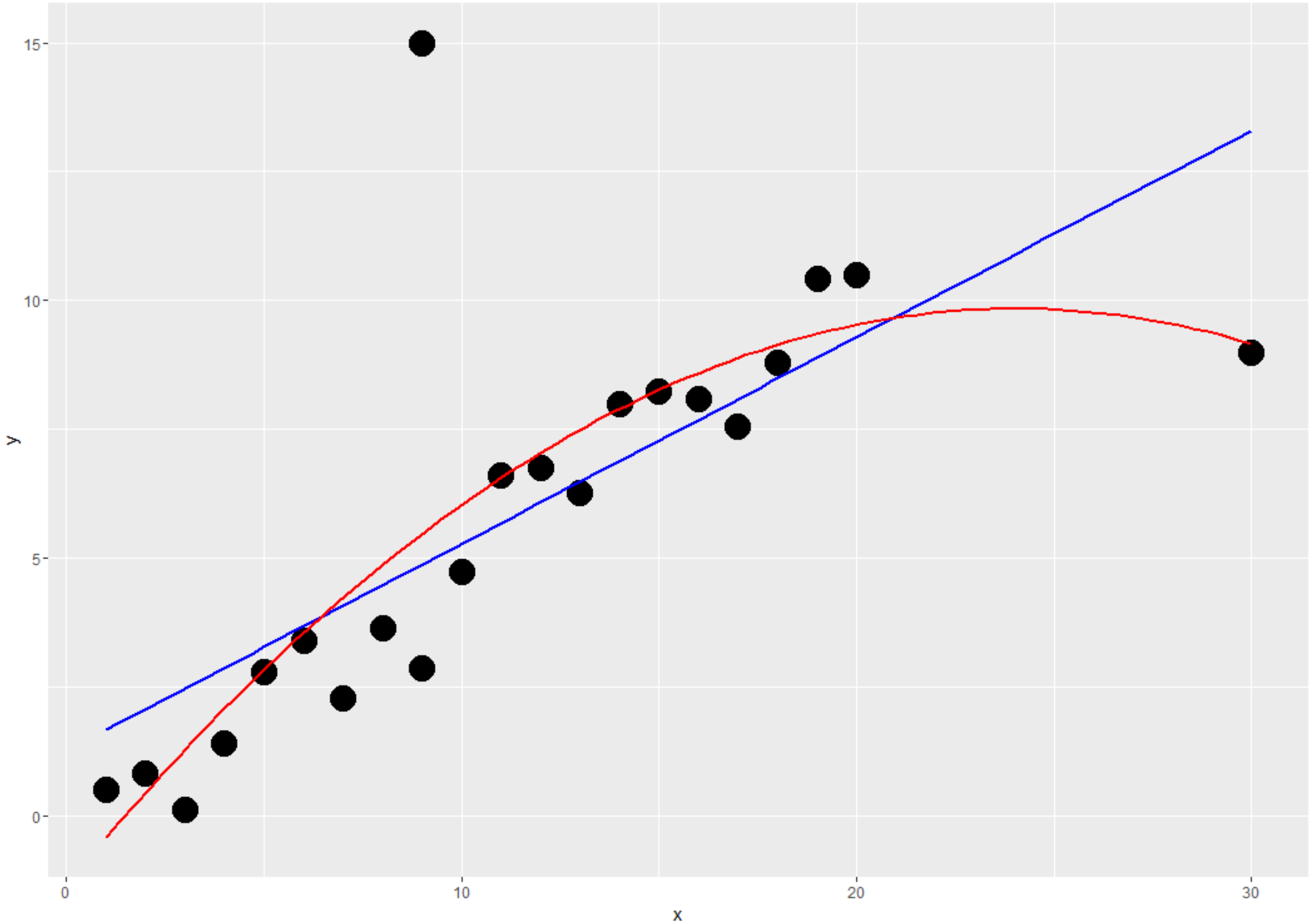
$$\mathbb{E}[Y] = \beta_0 + \beta_1(X) + \beta_2(X^2)$$

# Non-linear effects

| x | x^2 | x^3 |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 1 | 1 |
| 2 | 4 | 8 |
| 3 | 9 | 27 |
| 4 | 16 | 64 |
| 5 | 25 | 125 |
| 6 | 36 | 216 |
| 7 | 49 | 343 |
| 8 | 64 | 512 |
| 9 | 81 | 729 |
| 10 | 100 | 1000 |

# Non-linear effects

Problems with polynomials – the tails shift the whole function



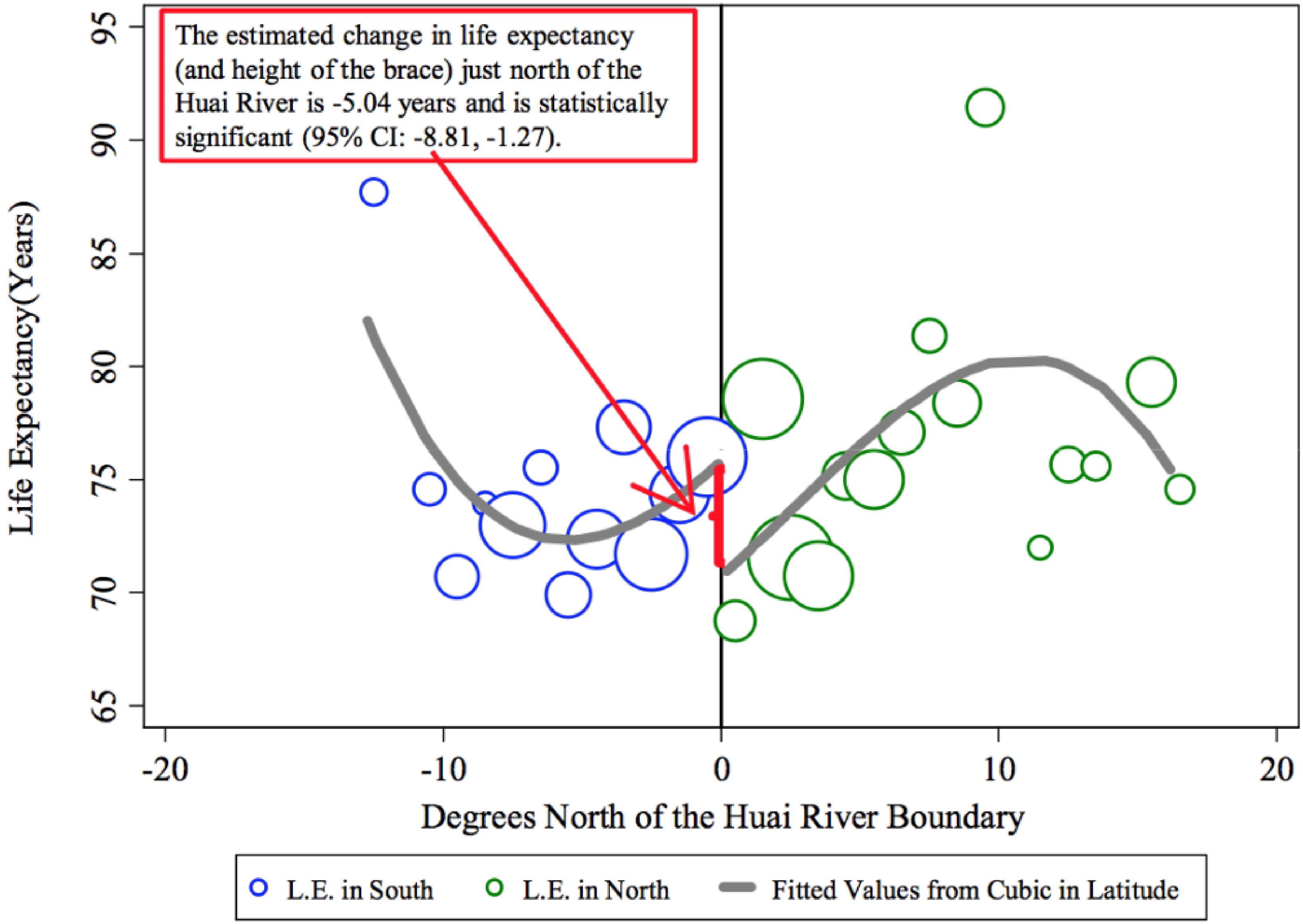



**Fig. 3.** The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

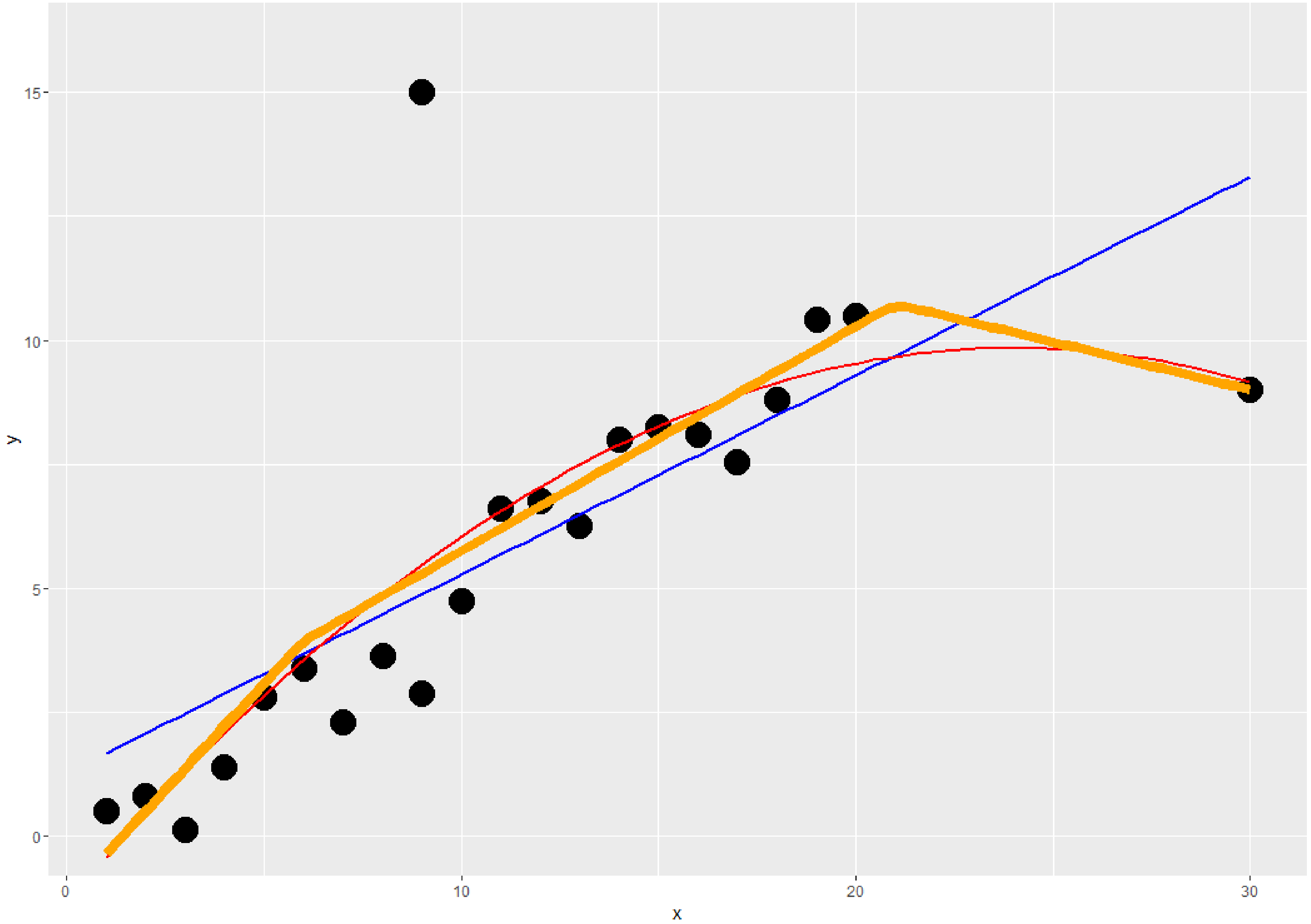
# Non-linear effects

Alternative – *splines*

Example *knot* at 5 (linear)

| x | b1 | b2 |
|---|---|---|
| 0 | 0.0 | 0.0 |
| 1 | 0.2 | 0.0 |
| 2 | 0.4 | 0.0 |
| 3 | 0.6 | 0.0 |
| 4 | 0.8 | 0.0 |
| 5 | 1.0 | 0.0 |
| 6 | 0.8 | 0.2 |
| 7 | 0.6 | 0.4 |
| 8 | 0.4 | 0.6 |
| 9 | 0.2 | 0.8 |
| 10 | 0.0 | 1.0 |

# Non-linear effects

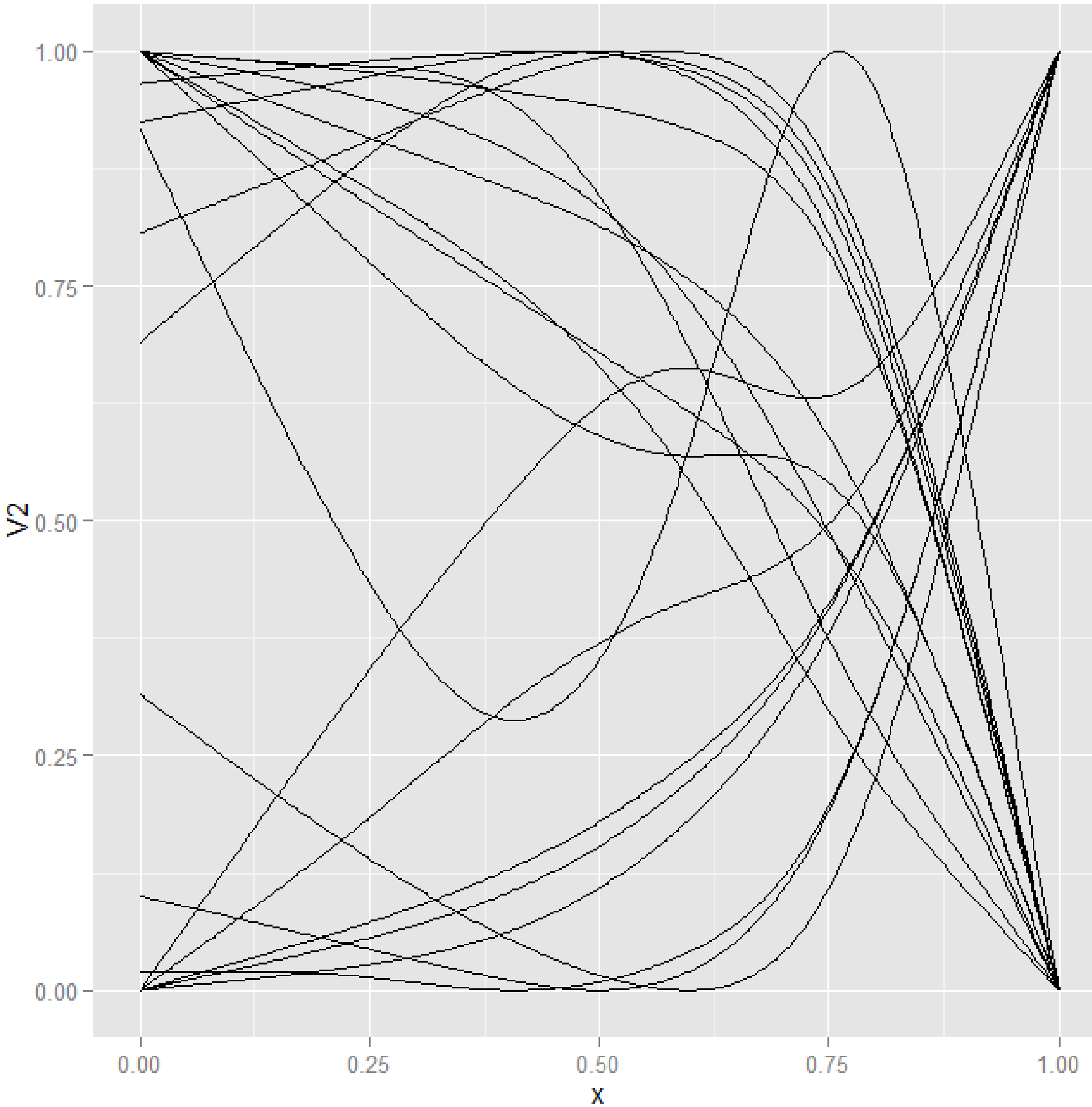


# Non-linear effects

- Why are splines better?
  - Tails have less of an effect on other parts of the function
  - Very few regularly placed knots
  - Function is still continuous

- See Harrell's *Regression Modelling Strategies* for the source of motivation for regression splines

# Non-linear effects



# Homework & Next Weeks Class

Lab Assignment

Use R, Stata or SPSS to compare effect estimates from linear and Poisson regression. Also gives an example of using restricted cubic splines

For Next Week – Propensity Score Matching

- Experimental and Quasi-Experimental, Chapters 4 & 5
- Apel and Sweeten. 2010. Propensity score matching in criminology and criminal justice
- Berk. 1983. An introduction to sample selection bias in sociological data.