Lecture 01
Introduction and Motivation

02 September 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Course overview

Linear Models is both a capstone to the 241/242 sequence and the breadth to compliment 610's depth. It also serves as a link between the statistical inference courses and the applied data analysis courses.

Linear Models is both a capstone to the 241/242 sequence and the breadth to compliment 610's depth. It also serves as a link between the statistical inference courses and the applied data analysis courses.

Topics will be oriented around linear models (obviously) but the course is somewhat of a hodgepodge of topics and applications.

From the course catalogue:

The geometry of least squares; distribution theory for normal errors; regression, analysis of variance, and designed experiments; numerical algorithms, with particular reference to the R statistical language.

My interpretation:

Three parts:

My interpretation:

Three parts:

1. Classical linear model theory

My interpretation:

Three parts:

1. Classical linear model theory
    - 1.1 Multivariate regression; normal equations and OLS
    - 1.2 Finite sample distribution theory
    - 1.3 Large sample theory
    - 1.4 Weighted least squares and model assumptions

My interpretation:

Three parts:

1. Classical linear model theory
   1.1 Multivariate regression; normal equations and OLS
   1.2 Finite sample distribution theory
   1.3 Large sample theory
   1.4 Weighted least squares and model assumptions
2. Computational techniques and penalized estimation

My interpretation:

Three parts:

1. Classical linear model theory
   1.1 Multivariate regression; normal equations and OLS
   1.2 Finite sample distribution theory
   1.3 Large sample theory
   1.4 Weighted least squares and model assumptions
2. Computational techniques and penalized estimation
   2.1 Solving least squares and sensitivity analysis
   2.2 Iterative methods for solving least squares
   2.3 Ridge and lasso regresion

My interpretation:

Three parts:

1. Classical linear model theory
   - 1.1 Multivariate regression; normal equations and OLS
   - 1.2 Finite sample distribution theory
   - 1.3 Large sample theory
   - 1.4 Weighted least squares and model assumptions
2. Computational techniques and penalized estimation
   - 2.1 Solving least squares and sensitivity analysis
   - 2.2 Iterative methods for solving least squares
   - 2.3 Ridge and lasso regresion
3. Additional topics

My interpretation:

Three parts:

1. Classical linear model theory
   1.1 Multivariate regression; normal equations and OLS
   1.2 Finite sample distribution theory
   1.3 Large sample theory
   1.4 Weighted least squares and model assumptions
2. Computational techniques and penalized estimation
   2.1 Solving least squares and sensitivity analysis
   2.2 Iterative methods for solving least squares
   2.3 Ridge and lasso regresion
3. Additional topics
   3.1 Bayesian regression
   3.2 Robust techniques
   3.3 GLMs

Class Survey

If $\{y_1, \ldots y_n\}$ are independent observations of a random variable distributed as $\mathcal{N}(\mu, \sigma^2)$, do you know how to calculate the maximum likelihood estimators of μ and σ^2?

Are you familiar with simple linear regression models?

$$y_i = \alpha + x_i \cdot \beta + \sigma \cdot \epsilon_i$$

Are you familiar with simple linear regression models?

$$y_i = \alpha + x_i \cdot \beta + \sigma \cdot \epsilon_i$$

Specifically, have you seen (don't need to remember) the ordinary least squares estimators for $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma^2}$.

Are you familiar with multivariate linear regression models?

$$y_i = \sum_j x_{i,j} \cdot \beta_j + \sigma \cdot \epsilon_i$$

And the associated (matrix form) of the estimators $\widehat{\beta}$ and $\widehat{\sigma^2}$?

Could you describe the properties that make a matrix D a *positive definite* matrix?

Are you familiar with the Cholesky decomposition of a matrix?

Are you familiar with the Cholesky decomposition of a matrix?

How about the QR or LU decomposion?

Have you computed by hand the Cholesky, QR, or LU decomposition of a matrix?

Have you used the lasso

$$\arg\min_b \left\{ ||y - Xb||_2^2 + \lambda \cdot ||b||_1 \right\}$$

Have you used the lasso

$$\arg\min_{b} \left\{ ||y - Xb||_2^2 + \lambda \cdot ||b||_1 \right\}$$

Or ridge regression

$$\arg\min_{b} \left\{ ||y - Xb||_2^2 + \lambda \cdot ||b||_2^2 \right\}?$$

Syllabus, ect.

Suggested Prerequisites:

Suggested Prerequisites:

- Linear Algebra at the level of MATH 222

Suggested Prerequisites:

- Linear Algebra at the level of MATH 222
- Statistical theory at the level of STAT 242

Suggested Prerequisites:

- Linear Algebra at the level of MATH 222
- Statistical theory at the level of STAT 242
- Some familiarity with a statistical software or programming language, preferably R

Grading

- 70% Problem Sets (10% each)

Grading

- 70% Problem Sets (10% each)
- 15% Mid-Term I (2015-10-12)

Grading

- 70% Problem Sets (10% each)
- 15% Mid-Term I (2015-10-12)
- 15% Mid-Term II (2015-11-18)

Problem Sets:
Problem sets are assigned roughly once every two weeks; this yields a total of 7 sets. You may discuss problem sets with other students, but must write up your own solutions. This means that you should have no need to look at other student's final written solutions.

Tentative due dates for problem sets: 09-14, 09-28, 10-05, 10-19, 11-02, 11-09 and 12-16. The final assignment is due the last day of reading period and may be handed in to the office at 24 Hillhouse.

STAT 312 vs. STAT 612

STAT 312 vs. STAT 612

Same requirements and assignments; final grades will be determined seperately.

STAT 312 vs. STAT 612

Same requirements and assignments; final grades will be determined seperately.

I strongly encourage undergraduates to have taken linear algebra and at least STAT 238 or STAT 242.

STAT 312 vs. STAT 612

Same requirements and assignments; final grades will be determined seperately.

I strongly encourage undergraduates to have taken linear algebra and at least STAT 238 or STAT 242.

I will teach with a graduate focus only in the sense that we will be concerned with **content** over **grades**.

Website

http://euler.stat.yale.edu/~tba3/stat612

Yale University STAT 312/6

euler.stat.yale.edu/~tba3/stat612/

Fork me on GitHub

STAT 312/612: Linear Models

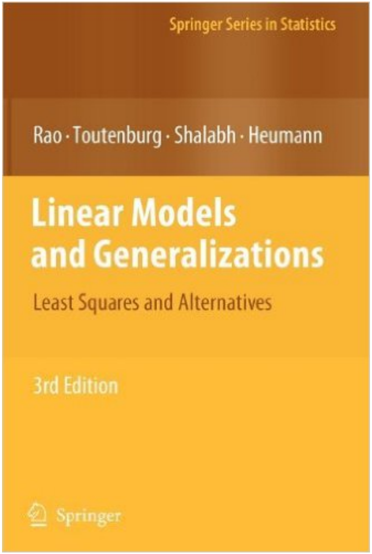
Course Notes and Assignments

Fall 2015
Monday, Wednesdays 11:35 - 12:50
17 Hillhouse, Rm 115

Instructor: Taylor Arnold
E-mail: taylor.arnold@yale.edu

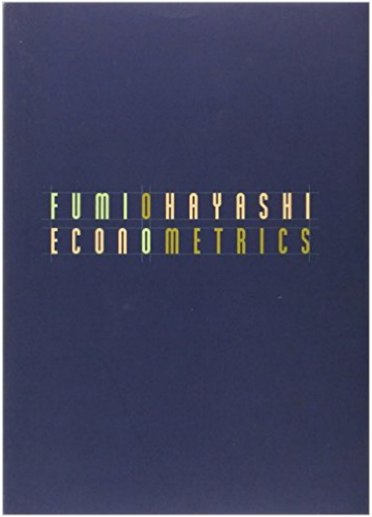
Date	Description	Resources	References
2015-09-02	Simple linear model assumptions and MLEs	[Syllabus] [Lecture 01]	RT 2.1-2.7
2015-09-07	Hypothesis tests; best linear unbiased estimators	**[Lecture 02]**	RT 2.8-2.10

Texts



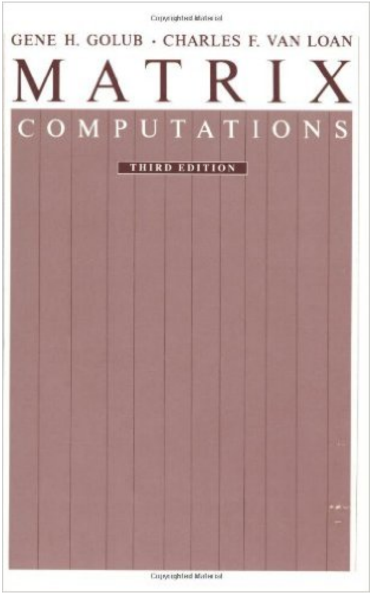
Rao, Calyampudi R., et al. *Linear models.* Springer New York, 2008.

- Available digitally through Springer Link (free pdfs from Yale network)
- Solid all-around reference on linear models
- Many special cases and extensions; will be a source of many problem set questions



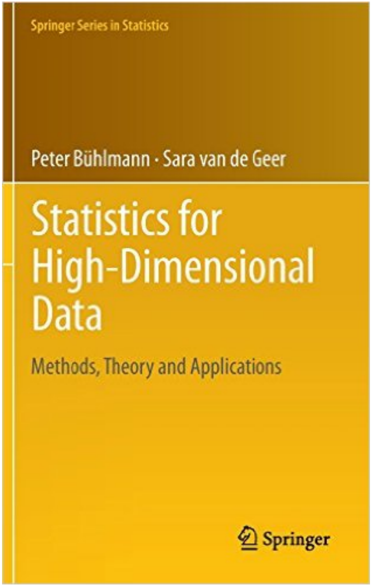
Hayashi, Fumio. *Econometrics.* Princeton University Press. (2000).

- My go-to reference for multivariate regression results and notation
- Intended for econometrics audience, but very thorough and theoretically sound
- Will primarily look at first two chapters only
- Focused on random design (stochastic X) and GMM methods
- Intro chapter available from publisher as a free pdf



Golub, Gene H., and Charles F. Van Loan. *Matrix computations.* Vol. 3. JHU Press, 2012.

- Considered the canonical reference on numerical linear algebra
- Not easily available online
- Will quickly go through the chapter on least squares estimators



Bühlmann, Peter, and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

- Available digitally through Springer Link (free pdfs from Yale network)
- A good reference for ℓ_1-penalized estimation
- Ignoring first 100 pages, gives a very thorough grounding on the basic theory and extensions
- Will reference this a lot when we study penalized estimators

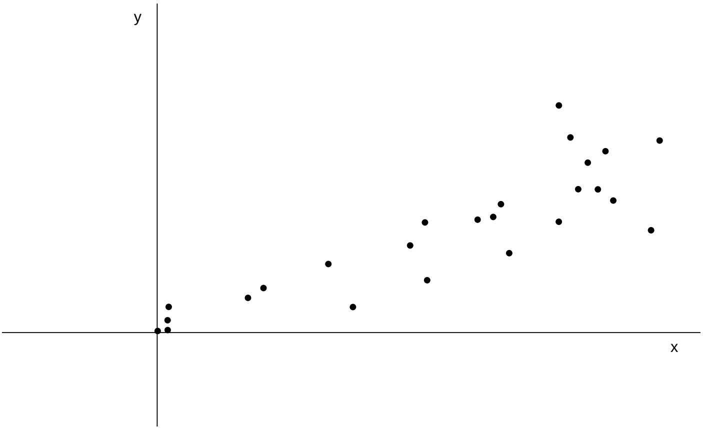
Me!

Joint appointment at Yale Statistics and AT&T Labs Research

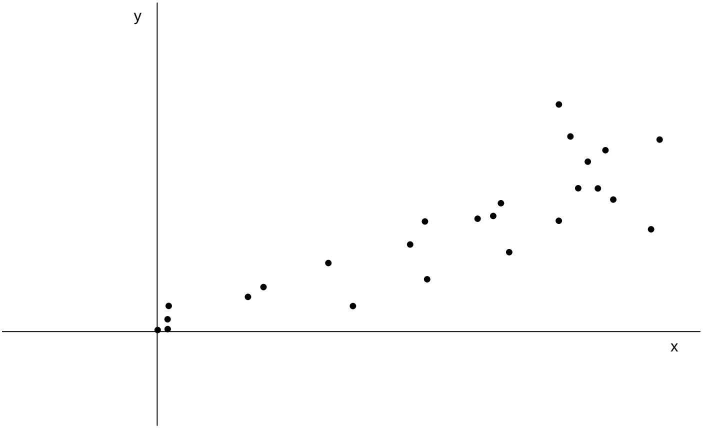
- Research focus on large-scale data analysis (think, petabytes)
- One focus is on encoding sparsity through penalized estimation
- Applications to humanities and social sciences through with analysis of image, text, and video corpora

What exactly are Linear Models?

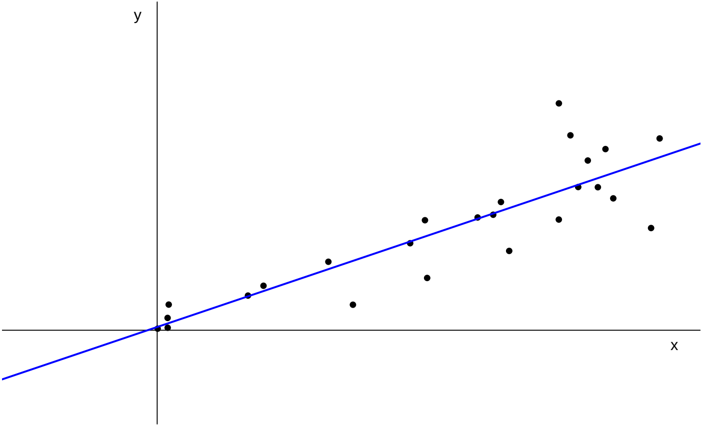
Consider observing pairs of points (x_i, y_i), which we can graphically represent by a scatter plot.



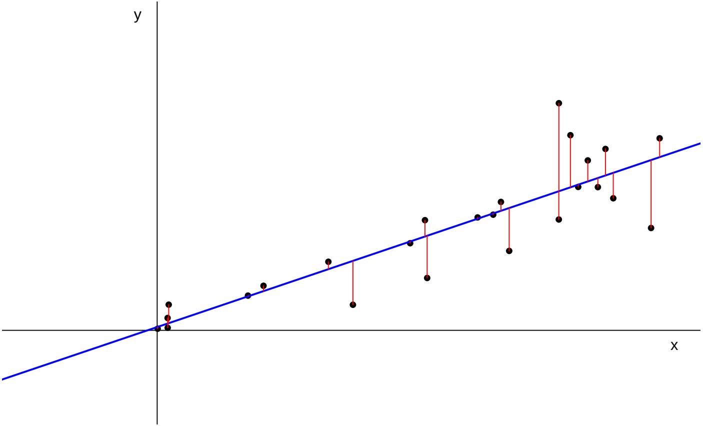
A **simple linear model** assumes that the mean of each y_i conditioned on x_i is a linear function of x_i.



Visually, we can think of this as a line through the data.



For a reasonable fit, the **residuals**, shown in red, should have a mean close to zero. The should also be ‘small’ in some sense.



Symbolically, the simple linear regression model assumes that:

$$\mathbb{E}(y_i|x_i) = \alpha + x_i \cdot \beta \tag{1}$$

The goal, typically, is to find point estimates and conduct inference on the unknown parameters α and β.

Classic examples of quantities modelled with simple linear regression:

- College GPA $\sim$ SAT scores
- Change in GDP $\sim$ change in unemployment
- House price $\sim$ number of bedrooms
- Species heart weight $\sim$ species body weight
- Fatilities per year $\sim$ speed limit

Classic examples of quantities modelled with simple linear regression:

- College GPA $\sim$ SAT scores
- Change in GDP $\sim$ change in unemployment
- House price $\sim$ number of bedrooms
- Species heart weight $\sim$ species body weight
- Fatilities per year $\sim$ speed limit

Notice that these simple linear regressions are simplifications of more complex relationships between the variables in question.

What sign would be expect of β (the slope) from each of these?

- College GPA $\sim$ SAT scores

What sign would be expect of β (the slope) from each of these?

- College GPA $\sim$ SAT scores $\quad \beta > 0$
- Change in GDP $\sim$ change in unemployment

What sign would be expect of β (the slope) from each of these?

- College GPA $\sim$ SAT scores $\beta > 0$
- Change in GDP $\sim$ change in unemployment $\beta < 0$
- House price $\sim$ number of bedrooms

What sign would be expect of β (the slope) from each of these?

- College GPA $\sim$ SAT scores $\beta > 0$
- Change in GDP $\sim$ change in unemployment $\beta < 0$
- House price $\sim$ number of bedrooms $\beta > 0$
- Species heart weight $\sim$ species body weight

What sign would be expect of β (the slope) from each of these?

- College GPA $\sim$ SAT scores $\beta > 0$
- Change in GDP $\sim$ change in unemployment $\beta < 0$
- House price $\sim$ number of bedrooms $\beta > 0$
- Species heart weight $\sim$ species body weight $\beta > 0$
- Fatilities per year $\sim$ speed limit

What sign would be expect of β (the slope) from each of these?

- College GPA $\sim$ SAT scores $\quad \beta > 0$
- Change in GDP $\sim$ change in unemployment $\quad \beta < 0$
- House price $\sim$ number of bedrooms $\quad \beta > 0$
- Species heart weight $\sim$ species body weight $\quad \beta > 0$
- Fatilities per year $\sim$ speed limit $\quad \beta < 0$

A **(general) linear model** is similiar to the simple variant, but with a multivariate $x \in \mathbb{R}^p$ and a mean given by a hyperplane in place of a single line.

$$\mathbb{E}(y_i|x_i) = \alpha + \sum_j x_{i,j} \cdot \beta_j \tag{2}$$

- General principles are the same as the simple case

- General principles are the same as the simple case
- Math is more difficult because we need to use matricies

- General principles are the same as the simple case
- Math is more difficult because we need to use matricies
- Interpretation is more difficult because the β_j are effects conditional on the other variables

For example, consider these two variable regressions:

- College GPA $\sim$ SAT scores, secondary school GPA
- Change in GDP $\sim$ change in unemployment, inflation
- House price $\sim$ number of bedrooms, area of the house
- Species heart weight $\sim$ species body weight, species height
- Fatilities per year $\sim$ speed limit, minimum legal speed

For example, consider these two variable regressions:

- College GPA $\sim$ SAT scores, secondary school GPA
- Change in GDP $\sim$ change in unemployment, inflation
- House price $\sim$ number of bedrooms, area of the house
- Species heart weight $\sim$ species body weight, species height
- Fatilities per year $\sim$ speed limit, minimum legal speed

Many would retain the same signs as the simple linear regression, but the magnitudes would be smaller. In some cases, it is possible for the relationship to flip directions when a second (highly correlated) variable is added.

What might be an explanation of the following signs:

- College GPA $\sim$ SAT scores, secondary school GPA

What might be an explanation of the following signs:

- College GPA $\sim$ SAT scores, secondary school GPA
- House price $\sim$ number of bedrooms, area of the house

What might be an explanation of the following signs:

- College GPA $\sim$ SAT scores, secondary school GPA
- House price $\sim$ number of bedrooms, area of the house
- Species heart weight $\sim$ species body weight, species height

Extensions to linear models include:

Extensions to linear models include:

- generalized linear models:

$$\mathbb{E}(y|x) = g^{-1}(x^t\beta)$$

Extensions to linear models include:

- generalized linear models:

$$\mathbb{E}(y|x) = g^{-1}(x^t\beta)$$

- additive models:

$$\mathbb{E}(y|x) = f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k)$$

Extensions to linear models include:

- generalized linear models:

$$\mathbb{E}(y|x) = g^{-1}(x^t\beta)$$

- additive models:

$$\mathbb{E}(y|x) = f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k)$$

- Generalized Additive Model for Location, Scale and Shape (GAMLSS):

$$\begin{aligned}
\mathbb{E}(y|x) &= f_{1,1}(x_1) + f_{1,1}(x_2) + \cdots + f_{1,1}(x_k) \\
\mathbb{E}(y^2|x) &= f_{1,2}(x_1) + f_{2,2}(x_2) + \cdots + f_{k,2}(x_k) \\
&\vdots \\
\mathbb{E}(y^q|x) &= f_{1,q}(x_1) + f_{2,q}(x_2) + \cdots + f_{k,q}(x_k)
\end{aligned}$$

Machine Learning & Linear Models

Machine learning is a closely related field to statistics; most researches that I know think of there being a spectrum of research between the two rather than a clear dividing line.

Machine learning is a closely related field to statistics; most researches that I know think of there being a spectrum of research between the two rather than a clear dividing line.

If forced to catagorize them, I would describe statistics as being primarily concerned with **inference** and machine learning with **prediction**.

Question

With powerful methods such as neural networks, support vector machines, and gradient boosted trees, is there space for linear models in machine learning?

Answer

Yes!

 # 1. When the number of parameters is close to or exceeds the number of observations, particularly if the data matrix X is sparse.

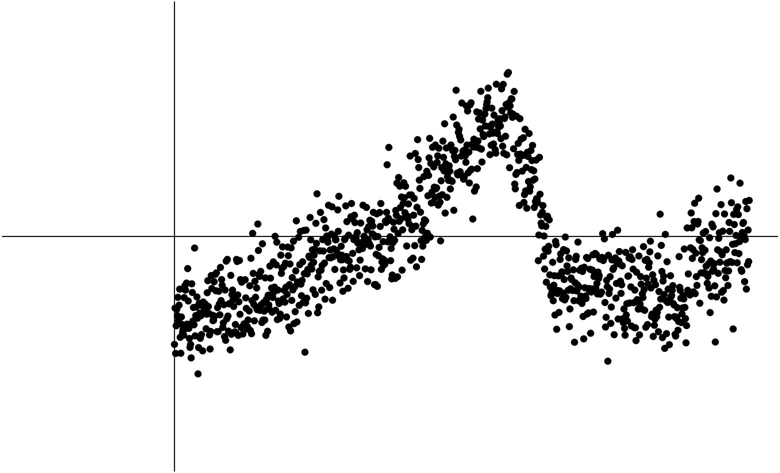
2. Creating meta-variables as an input to other ML techniques or to blend the outputs from ensemble learning.

 # 3. Working with data that have difficult to work with distributions, such as quantile regression on heavy-tailed errors (i.e., Cauchy, Lévy).

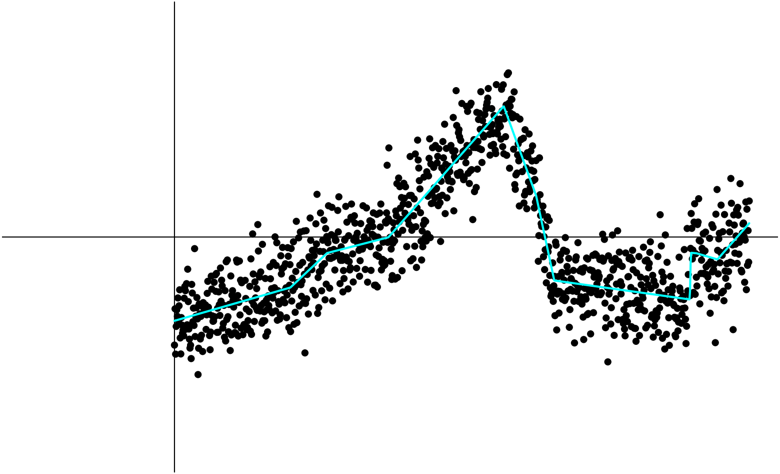
4. Projecting into high dimensional spaces (where often we have more predictors that observations and spare data matrices).

When 'Linear' Isn't

Consider the following set of data points (x_i, y_i). The relationship between x and y is highly non-linear.



The true mean (from which I simulated) is given by:



It would at first seem that we can't model this response with a linear model. However, that is not the case because it is only β that needs to be linear, not the x values.

It would at first seem that we can't model this response with a linear model. However, that is not the case because it is only β that needs to be linear, not the x values.

For example, the following is a linear model:

$$\mathbb{E}(y|x) = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3$$

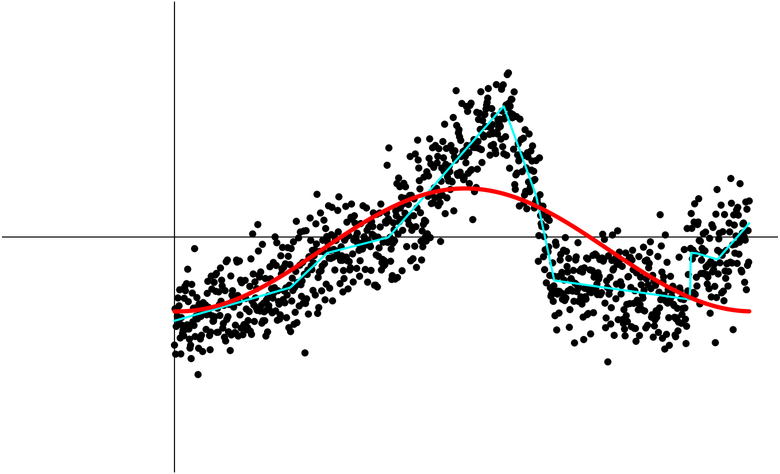
Which will fit a polynomial to the data.

An alternative that works better here, is a Fourier basis:

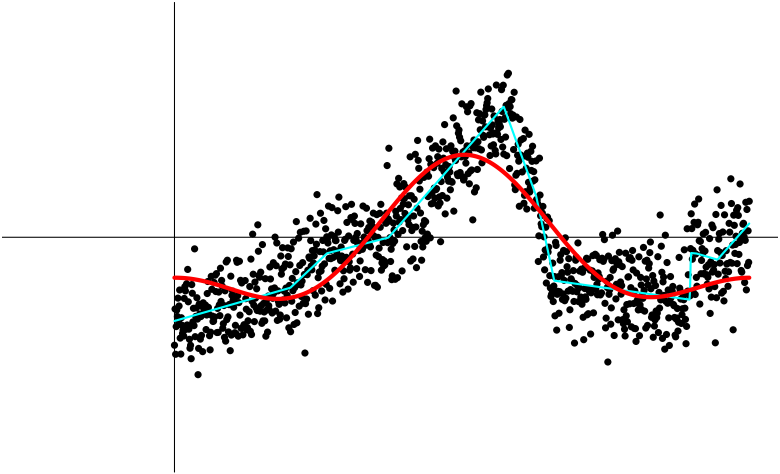
$$\mathbb{E}(y|x) = \beta_0 + \sum_{j=1}^{k} \beta_j \cos(j * x) + \sum_{j=1}^{k} \beta_{k+j} \sin(j * x)$$

And we can adjust the fit appropriately by specifying the order k.

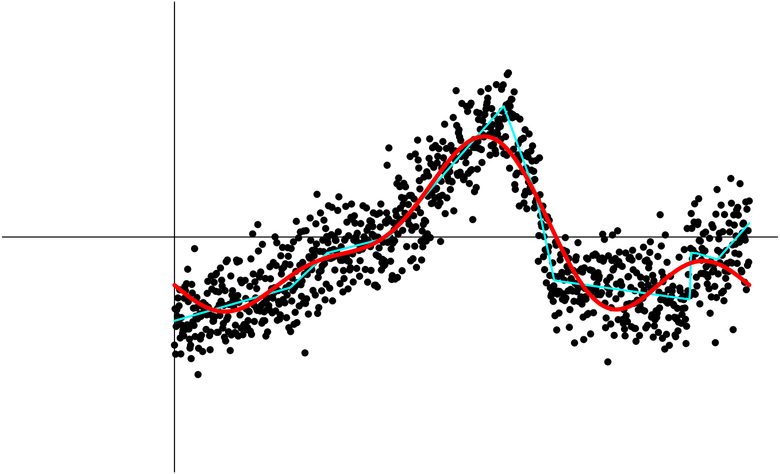
$k = 1$



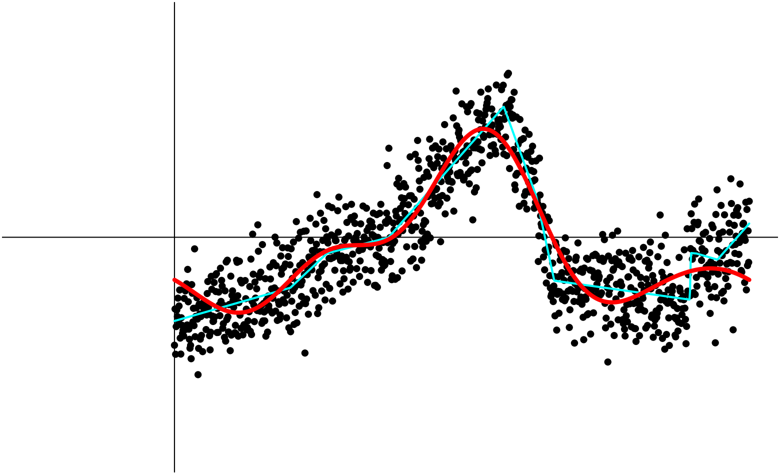
$k = 2$



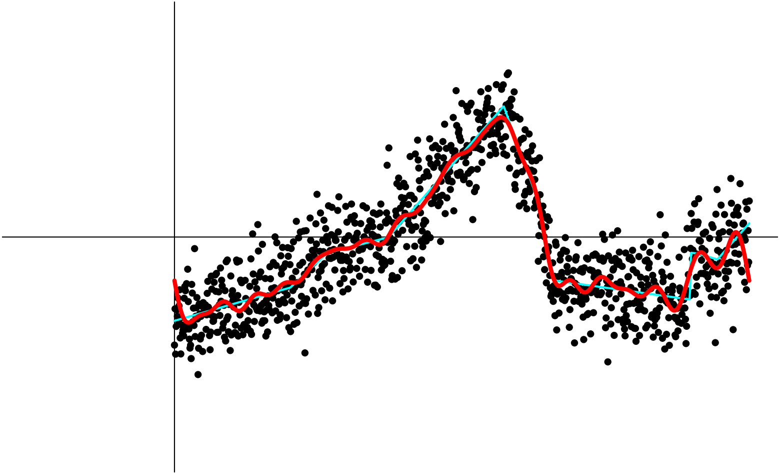
$k = 3$



$k = 4$



$k = 20$



Questions, thoughts or concerns?

Questions, thoughts or concerns?

website: `http://euler.stat.yale.edu/~tba3/stat612`
e-mail: `taylor.arnold@yale.edu`