Basic Linear Regression

Richard A. Reitmeyer
August 2016

Objectives

- Participants should be able to:
  - Create a basic linear regression in R
  - Compare models
  - Evolve simple models into more sophisticated ones by hand
  - Explain the basic linear algebra behind linear regression

Topics

- “Doing” data science
- Linear models
- Linear algebra
- Model matrix
- Titanic, in gory detail

Why is it called data “science”

- Our goal is data “science”
- Extract knowledge or inference from data
  - Easier than ever: Computers + Data
- Want to be able to predict things
  - Or classify things
  - Or infer cause (A/B testing)
- Goal: a **model** that lets us predict a **response** from **features**

“Doing” Data Science, simplified

- Start with a question
- Look at the data
- Build a simple model, with a simple modeling technique
- Extend the model

“Doing” Data Science, simplified

- Start with a question
- Look at the data
- Build a simple model, with a simple modeling technique
- Extend the model

These slides talk about one modeling technique and will only touch on other topics

Linear (Least Squares) Regression

- THE classic technique of data analysis
  - Used by Gauss around 1800 to predict Ceres' orbit
  - May go back a bit further to Legendre
  - Proven “optimal” by Gauss in early 1820s
- Requirements
  - Errors are Independent and Identically Distributed (IID)
    - Errors have zero mean
    - Errors are “homoscedastic” – independent of the value of prediction terms
  - Errors are normally distributed
- Assumption: prediction terms are known exactly

Sample Problems

- Titanic: predict who lived / died
  - Given training data of ~700 passengers, predict survival of another ~200
- Property Values
- School Performance

Theory

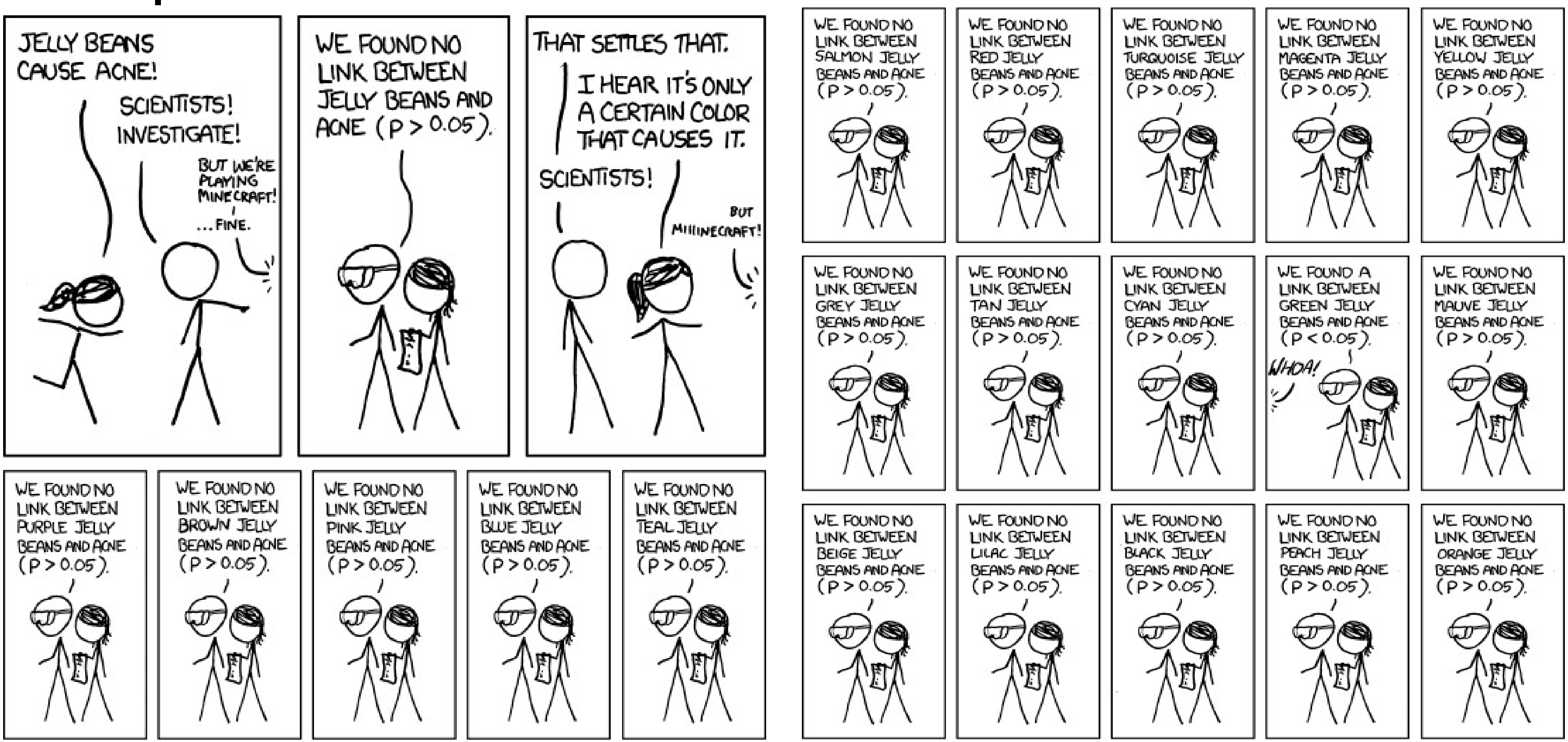
- All “science” starts with a question, and a theory about that question
- Your plausible theory could be implausible to me:
  - Men are stronger and better swimmers, so men have better odds than women
  - Edwardian era notions of stoic manliness and women's need of protection mean special treatment for women, so women have better odds
  - Women have more body fat, and so should survive better in cold water
- Data can eliminate testable theories that are “bad”

How confident?

- Before acting on a prediction, want to be know how confident we should be in it
  - How strongly should we act on the prediction?
- Use statistics to estimate likelihood
  - 95% likely: Academic, small sample benchmark
  - 99% likely, 99.9% likely: larger data sets, more need for certainty
- Won't cover model validation tonight, but it's an important topic

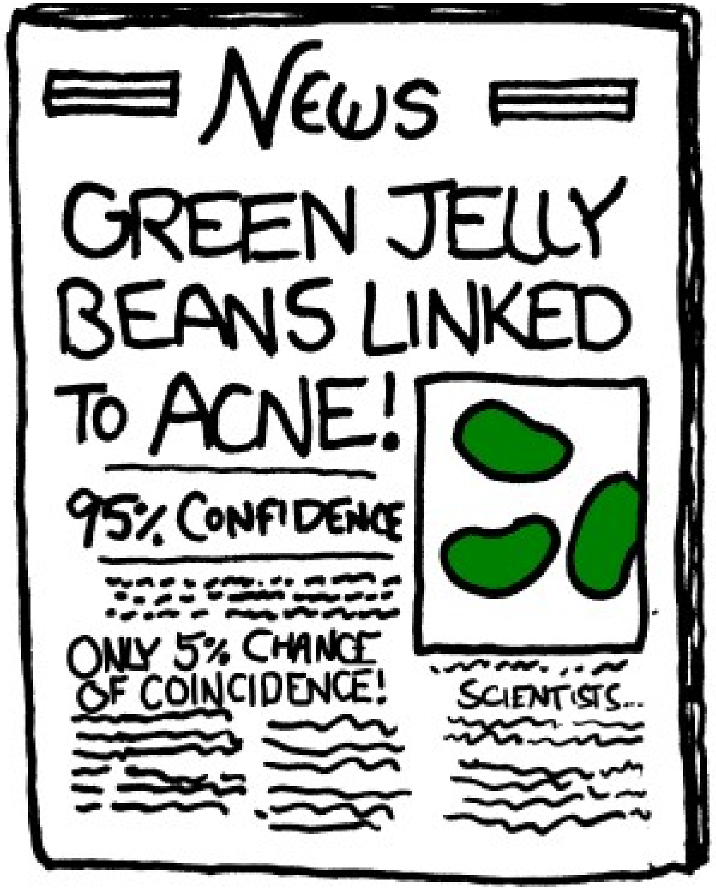
95% Confidence

- https://xkcd.com/882/



95% Confidence

- https://xkcd.com/882/



Linear Models

- A linear model is one where the response is related to the predictors in strictly additive way
  - Remember algebraic geometry: $y = m*x + b$?
- Linear model is $y = b0 + b1*x1 + b2*x2 + \ldots + bn*xn + e$
  - y is response, the thing to predict
  - x is a predictor term, a known feature or something derived from a feature (or features)
  - b is a coefficient
  - e is error --- must be gaussian, IID!

Simple or Complex

- Linear models can be almost as simple or complex as you like:
  - Dinner_bill = baseline_taxi_fare + avg_food_cost*diners + avg_beer_cost*drinkers
  - Mpg = baseline + b1*(1/weight)+b2*(1/weight^2) + b3*(1/horsepower)
  - Home_price = baseline + b1*sqft + b2*neighborhood + b3*sqft:neighborhood
- Counterexample:
  - mpg = baseline + (1/hp)^b1
- There are also extensions to linear models (“generalized”) to handle some non-gaussian distributions, used for classification, but we'll ignore those in this talk.

Three Minutes of Linear Algebra

- Way to write, work with, and solve large number of large linear equations

4 * b1	+	3 * b2	+	5 * b3	+	1 * b4		15
9 * b1	+	1 * b2	+	3 * b3	+	6 * b4	=	12
2 * b1	+	2 * b2	+	2 * b3	+	2 * b4		20
7 * b1	+	7 * b2	+	1 * b3	+	1 * b4		34

$$\begin{bmatrix} 4 & 3 & 5 & 1 \\ 9 & 1 & 3 & 6 \\ 2 & 2 & 2 & 2 \\ 7 & 7 & 1 & 1 \end{bmatrix} \begin{bmatrix} b1 \\ b2 \\ b3 \\ b4 \end{bmatrix} = \begin{bmatrix} 15 \\ 12 \\ 20 \\ 34 \end{bmatrix}$$

$$\boxed{\mathbf{X}\, b = y}$$

Three Minutes of Linear Algebra

- Addition: **S** = **A** + **B**, element Sij is Aij + Bij
- Multiplication: **P** = **A B**, Pij is sum over all m of Aim * Bmj. Note **A B** != **B A** (not commutative)!
- There is a "transpose" operation, written A' or x', that swaps rows and columns:

4	3	5	1
9	1	3	6
2	2	2	2
7	7	1	1

'

=

4	9	2	7
3	1	2	7
5	3	2	1
1	6	2	1

x1
x2
x3
x4

'

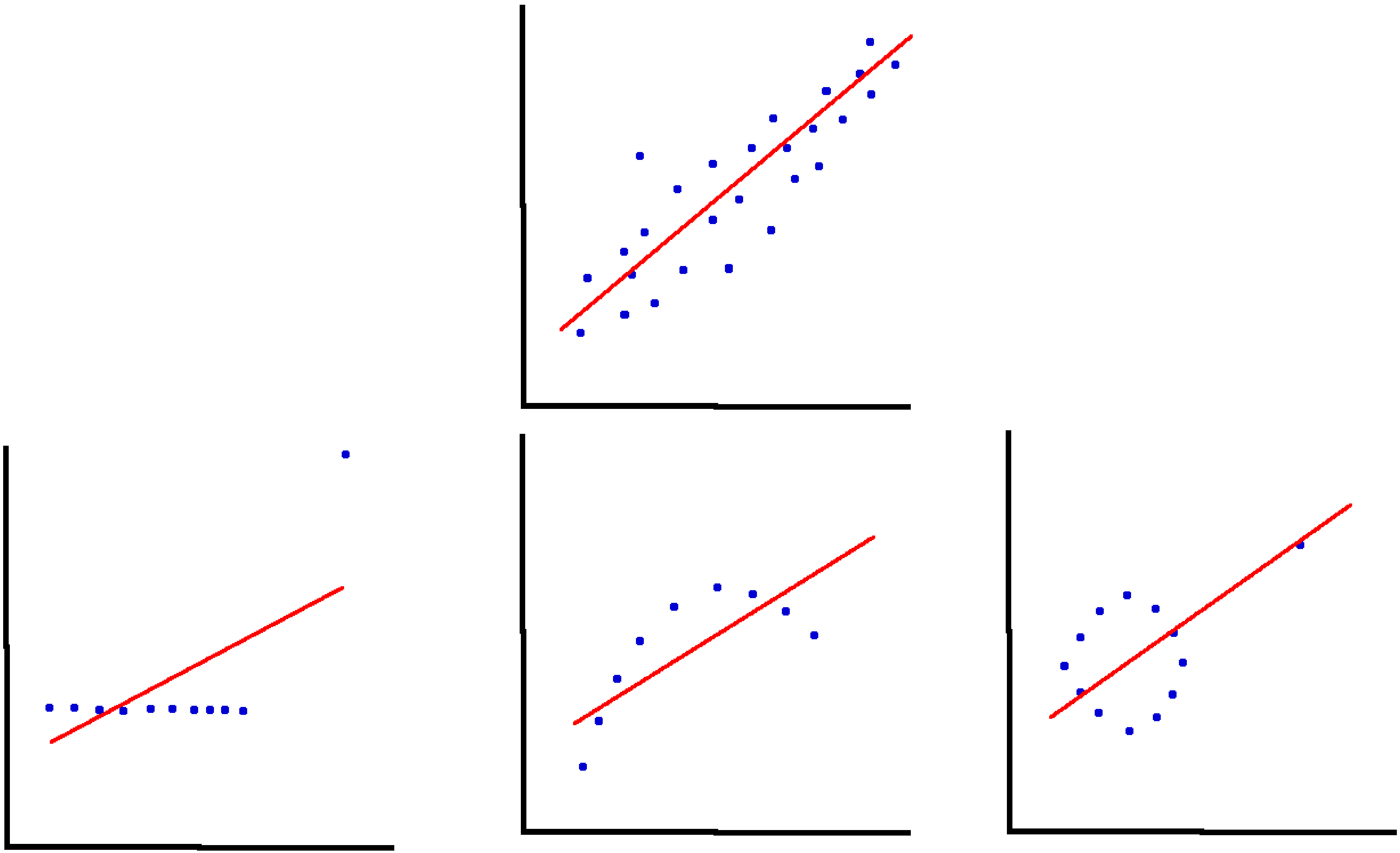
=

x1	x2	x3	x4

Three Minutes of Linear Algebra

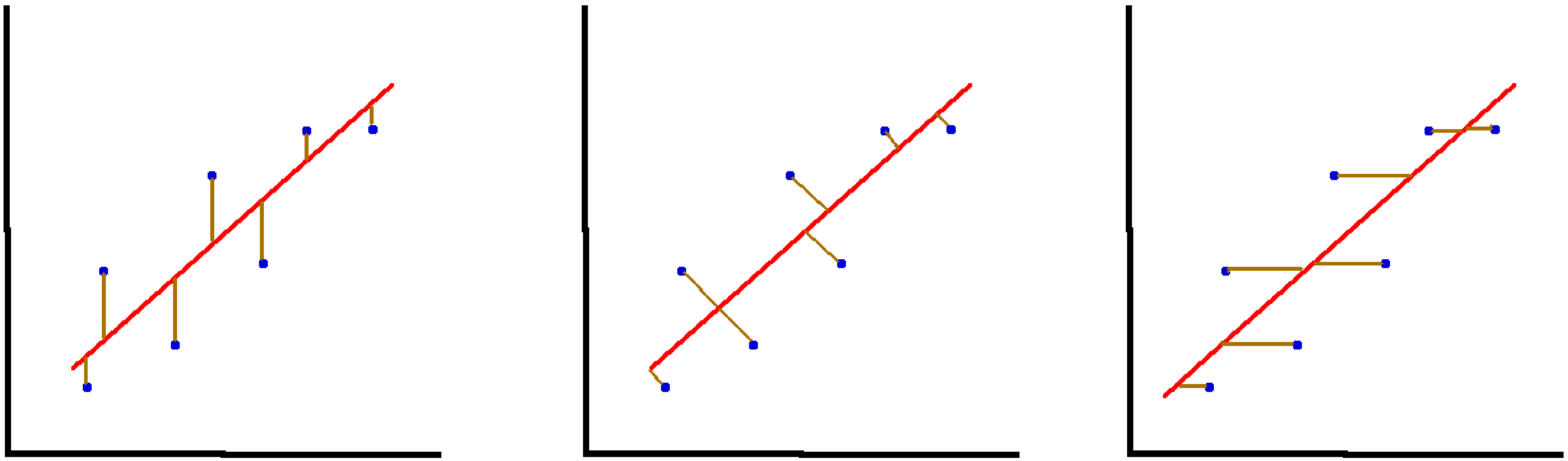
- Identity matrix: has 1s on the top-left/bottom right diagonal and zeros elsewhere
  - When anything is multiplied by identity, either side, get same matrix
- “Inverse” operation: defined as producing the matrix such that multiplying a matrix by its inverse (on either side) yields the identity matrix.
  - Similar to how the ordinary algebraic multiplicative inverse of n is 1/n, so 1/n*n = 1 and n*1/n = 1
  - Only square matrix can have inverse
  - Even square matrix can sometimes have no inverse (“ill-conditioned”), but “most” do have one

What is a Linear Fit?



Reminder: What is Minimized?

- residuals: actual – predicted
- errors: predicted - actual



Quality Metrics

- Many measures for “goodness” of a model
  - Might be a good topic for a lecture of its own
- Some important ones:

name	range	meaning
R^2	0..1	Explained variance, 0=bad, 1=perfect, so models with bigger R^2 better.
P	> 0	Probability model (or coefficient) is worthless, so models (coefficients) with smaller P better
confidence interval	N/A	Coefficient is likely to fall in this range
AIC/BIC	> 0	“Information loss” when comparing models, smaller better