Regression 3 Ways

Oliver Will

DataPhilly Workshop

2020-02-18

https://github.com/owill/DataPhilly20210218

About me

- PhD in applied mathematics from University of Southern California
- Post doc in the statistics department of the University of Washington
- Market researcher for 15 years
- Currently at Kantar
- Previously at IQVIA
- Interest in programming languages

Outline of the presentation

1. Refresher on linear regression
2. Do the same simulation, visualization, regression, and analysis 3 times in
   a. R
   b. Python
   c. Julia
3. Might not get to logistic regression

Beginner workshop, but we're going to go fast

Three statistical programming languages

R

- 1995
  - Graduate class in S-PLUS in 1997 and learned about R
- Scheme with Fortran subroutines
- Open source version of S-PLUS

Python

- 1991
  - Programming as a postdoc in 2001. Tried Perl and Python
- C scripting language

Julia

- 2009
  - Learned it for this presentation out of curiosity
- ?, but Hadoop is on the scene

Try to stay away from a discussion of which language is best

Why Julia?

Today on LinkedIn

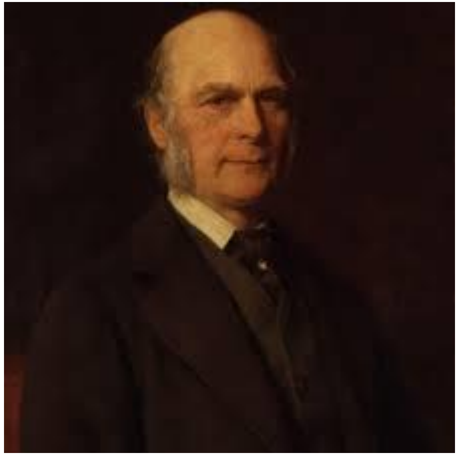
- 2,344 Python jobs in Philadelphia
  - 131 Python data scientist jobs
- 1,798 R jobs
- 12 Julia jobs

Look at this article

https://www.hpcwire.com/2020/01/14/julia-programmings-dramatic-rise-in-hpc-and-elsewhere/

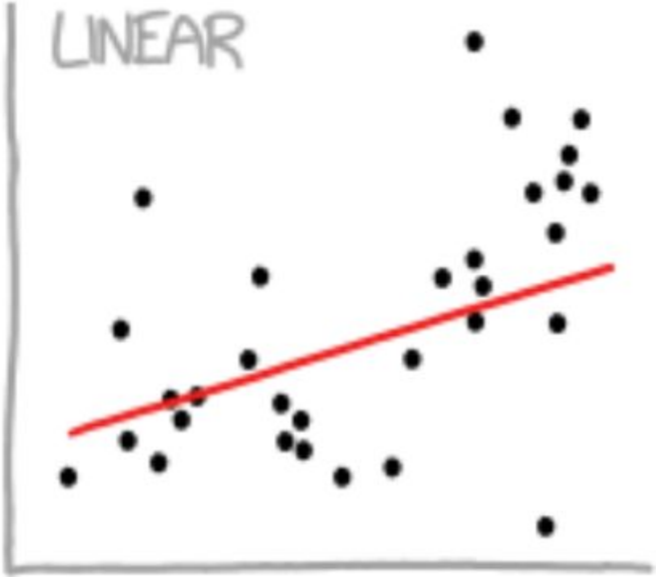
Might be a great language to mine Ethereum in

Linear regression refresher



Francis Galton, 1822-1911

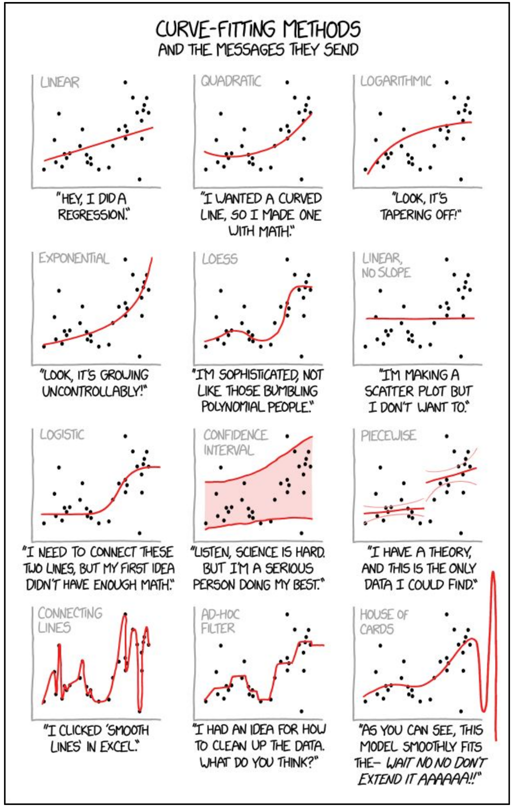
First linear regression



Karl Pearson, 1857-1936

Correlation

Xkcd: Curve-Fitting comic


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Randall Monroe
https://xkcd.com/2048/

Refresher on the theory of linear regression

- One variable called the response

$$y$$

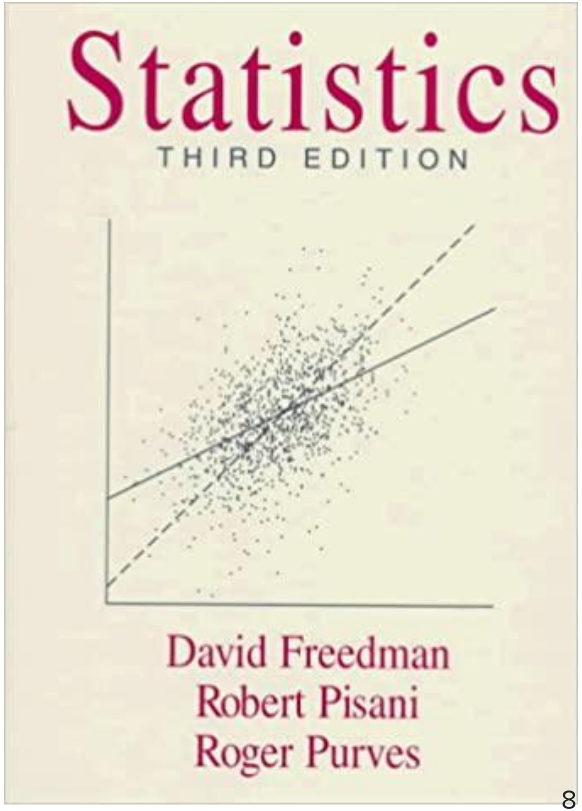
- One or more called predictors

$$x_1, \ldots, x_k$$

- One or more coefficients

$$\beta_0, \ldots, \beta_k$$

All the numbers are continuous, -∞ to +∞



Data for a linear regression

- *n* observations
- *n* responses, *y*
- *k+1* coefficients, β
- *n x (k+1)* predictors, *x* (and the ones!)

$$\boxed{y_1},\ \beta_0 + \beta_1 \boxed{x_{11}} + \ldots + \beta_k \boxed{x_{1k}}$$

$$\vdots$$

$$y_n,\ \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk}$$

See how the data are represented in the 3 languages

Desirable properties of a relationship of *x* to *y* through β

Follow Freedman, Pisani, and Purves (one predictor *k* = 1)

1. Mean of *x* be at the mean of *y* $\quad \bar{x} \quad \bar{y}$ are on the line

2. Spread of *x* maps to the spread of *y* $\quad SD_y / SD_x$

3. Change of *x* leads to what average change in *y*

$$\frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

How do we get these desirable properties?

Find β that minimizes the sum of squares between the predictors and response. Ordinary least squares (OLS).

$$\sum_{1 \leqslant i \leqslant n} \left(y_i - \beta_0 - \beta_1 x_{i1} \right)^2$$

Vector calculus problem. You get this solution

$$\widehat{\beta}_1 = cor(x, y) \frac{SD_y}{SD_x} \qquad \widehat{\beta}_0 = \widehat{\beta}_1 \bar{x} - \bar{y}$$

What are the key assumptions of linear regression?

LINE

L - Linearity

I - Independence of errors

N - Normality

E - Equality of errors

$$\varepsilon = y - \beta_0 - \beta_1 x$$

Gelman - Statistical Modeling, Causal Inference, and Social Science

0. Validity. Data should answer research question
1. Linearity and additivity
2. Independence of errors
3. Equal variance of errors
4. Normality of errors

Further assumptions if looking for causality

If you are concerned 3 and 4, you should do a hold-out sample as well

https://statmodeling.stat.columbia.edu/2013/08/04/19470/

Assumptions for linear regression

- Assumptions by Gelman are presented in order of importance
- 0 feels like common sense
- Assumptions 1 and 2 are the most important for using linear regression
  - Reason for introducing desirable properties
- Assumptions 3 and 4 are mainly for individual value prediction
  - I create a lot regressions and almost never check 3 and 4
- R^2 is a useful and not used for model checking

Maybe linear regression should be used as a descriptive statistic?

Continue in JupyterLab. . .

Observation

R

Amazed at the variety of syntax

My speed at getting things done

Python

Disappointed to extract data from Pandas into Scikit-Learn

Focusing on the object type leads the way

Julia

Impressed by the consistency

Only one with most recent version

Wonder what Spark.jl is like?

Surprised at how much I liked Jupyter Notebooks. I probably would remove Anacondas from my workflow if I were using Python or Julia more

Polls

If this were a live presentation, I’d be asking you these polls

Where are you in your career?

- Academic
- Private industry
- Volunteer / hobby / interest
- Student

Which statistical programming language do you use?

- R
- Python
- Julia
- Spark or Hadoop environment (including Java and Scala)
- Closed source statistical (SAS, Stata, Eviews, etc.)
- Mathematical/numerical (Mathematica, Matlab, Octave, etc.)
- Excel
- Other