Math 455 Class 1 January 22

Course website (with notes and syllabus)

people.math.binghamton.edu/dikran/455

Textbook: "Linear Models with R"

<u>SECOND EDITION</u> by Julian Faraway.

⚠ Warning: the theory in this textbook is very condensed, to focus on the practical.

Q: I want to code in Python, not R.

A: The textbook has been translated into Python:
"Linear Models with Python" by Julian Faraway.
(no 2nd ed. yet)

Q: My AI will do my homework for me.

A: You must be able to explain the AI's answers.
You will be asked to do this on tests.

Why is Regression Analysis?

Should the FDA approve XYZ blood pressure medication?

Does eating yogurt reduce the risk of cancer?

Are the police discriminating against black/minority/gay citizens?

Do ~~empoty~~ employers pay women less for the same work?

Do financial markets correctly price risk?

The way we answer these questions is: gather data and do regression analysis.

To understand what's going on in any of these debates, understand regression analysis.

Examples: Claudia Goldin (Nobel Econ '23)

Roland Fryer (John Bates Clark Medal)

What is Regression Analysis?

A: Drawing a line through a cloud of points.

Q: I could do that!

A: The line drawn is "best" under the assumptions of the analysis, provably so.

Also, we can "draw" a 9-dimensional plane in a 10-dimensional space.

Q: Wait! Where is this 10-dim'l space coming from?

A: Suppose we are gathering data on patients: height, weight, age, liver function blood test, etc. each of these variables gives a coordinate and if we measure 10 quantities, each patient is then associated to a point in a 10-dim space.

- this is the "variable space".

Suppose we have 100 patients.

We have a table of data : 100 rows and 10 cols.

each variable (say age) is associated to a column of 100 numbers : a point in $\mathbb{R}^{100}$. (the "subject space".)

Regression analysis is an orthogonal projection in the subject space.

First we should think about 2 dimensions:

We have a table of points in the plane: (x_i, y_i) points.

Y	X	Const.
y_1	x_1	1
y_2	x_2	1
$\vdots$	$\vdots$	$\vdots$
y_n	x_n	1

We want to find the "best" line $y = \beta_0 + \beta_1 x$ through these points.

Express $Y = \beta_0 (\text{Const}) + \beta_1 X + \text{err.}$

Thee "best" choice is the orthogonal projection of Y onto the plane spanned by X and Const.

orthogonal projection is distance - minimizing.

the "best" choices $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfy.

$$\text{dist}\left(Y, \hat{\beta}_0(\text{Const}) + \hat{\beta}_1 X\right) \text{ is minimal.}$$

$$\text{dist}^2\left(Y, \hat{\beta}_0(\text{Const}), \hat{\beta}_1 X\right)$$

$$= \text{dist}^2\left((y_1, \dots, y_n), (\hat{\beta}_0 + \hat{\beta}_1 x_1, \dots, \hat{\beta}_0 + \hat{\beta}_1 x_n)\right)$$

$$\left(y - (\hat{\beta}_0 + \hat{\beta}_1 x_1)\right)^2 + \dots + \left(y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)\right)^2$$

We are choosing $\hat{\beta}_0, \hat{\beta}_1$ to minimize the sum of the squares.

Regression Analysis is also called "least-squares".

Important: First HW is already up in Gradescope.

NOTE: Textbook has a website with scripts for all R code.