Data Science for Everyone

Week 13: Regression

Angela Lai

May 1, 2020

New York University

Outline

- Logistics
- Concept review
- Demo

Logistics

- **Optional** lab 10 due at 8 p.m. ET on Wednesday, May 6
- **Extended deadline**: Project due at 8 p.m. ET on Friday, May 8
- Final exam on Monday, May 11

Project Reminders

- Clearly indicate every question you answer
- Best practices: create a Jupyter Notebook like the homework assignments. Fill that out with your code and analyses and, in the end, print to PDF. Make a cell (or cells) for every question.
- Be very careful about how you interpret the regression coefficients, p-values, and confidence intervals
- Save as HTML, print to PDF, and submit on Gradescope

Concept Review

Key points:

- Is correlation symmetric? What about regression?

Concept Review

Key points:

- Is correlation symmetric? What about regression?
- Correlation is for linear association. Regression is about prediction and relationship
- Recall from last week's lab: for one-variable linear regression, $slope = r \cdot \left(\frac{s_y}{s_x}\right)$, where r is correlation and s_y and s_x are the standard deviations of y and x, respectively.
- In this class, your hypotheses should not contain the word "correlation"

Regression Review

We want to find a linear model for the relationship between x and y. That is, we want to find slope b and intercept a for $y = bx + a$.

Regression Review

Linear regression: a linear approach to modeling the relationship between a dependent variable and one or more independent variables.

The **correlation coefficient** measures linear association between two variables. We write it as r.

R^2: the proportion of the variance in the outcome (y) that our linear regression explains. In this context (one-variable least squares), $R^2 = r^2$.

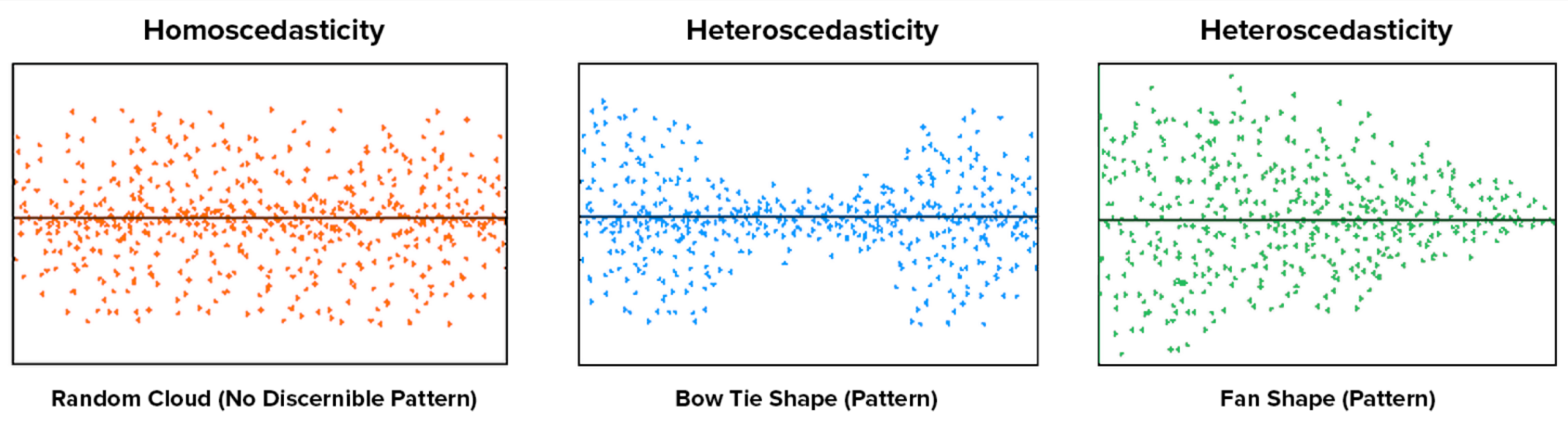
Review

Residual: The difference between the actual and predicted value of a variable. *true − predicted*

A collection of random variables is **homoscedastic** if all of its random variables have the same finite variance, a.k.a. homogeneity of variance.

A collection of random variables is **heteroscedastic** if there are sub-populations that have different variabilities from others.

Heteroscedasticity vs. homoscedasticity



(from last year's lecture slides)

Demo

Let’s see how we can do this with code!