

# Introduction to Linear Regression

GR 5205 / GU 4205  
Section 2/ Section 3

Columbia University  
Xiaofei Shi





# Course setup

## Basic administrative details:

Course website      [https://innerpeas.github.io/LRM\\_index.html](https://innerpeas.github.io/LRM_index.html)

- Instructor: Xiaofei Shi [xs2427@columbia.edu](mailto:xs2427@columbia.edu) OH: Tuesday 1:30 - 2:30
- TA: Navid Ardeshir [na2844@columbia.edu](mailto:na2844@columbia.edu) OH: Monday 11:00 - 12:00  
Daran Xu [dx2207@columbia.edu](mailto:dx2207@columbia.edu) OH: Monday 3:00 - 4:00

Office hours are on Zoom: 969 8687 3013

- Textbook: Kutner, Nachtsheim, Neter: *Applied Linear Regression Models*
- We will use Courseworks to post lecture notes and homework
- We will use Piazza for announcements and discussions:

GU 5205/GU 4205 Linear Regression Models



# Tentative Evaluation Plan

20% Final + 30% Midterm + 50%  $\max\{\text{homework average, exam average}\}$

- Homework: 5 homework in total;
- Midterm: October 27th & November 29th, in class;
- Final: See school schedule;
- Participation: Piazza, recitations, class survey.



# Class participation

- In person class only, but subject to changes
- According to the current return to school procedure, all class will be in person. In order to protect everyone in the classroom while provide the education we promised you,
  - Please wear a facial mask that covers your nose and mouth
  - Please keep 6-feet social distance
  - Follow the gateway testing



# Prerequisites

**Assume working knowledge of/proficiency with:**

- Probability
- Linear Algebra
- Statistics
- Programming (Python, R or Matlab)
- Formal mathematical thinking

If you fall short on any one of these things, it's certainly possible to catch up; but don't hesitate to talk to us.



# Goal of this course

- Data analysis
- R, Python, Matlab, Mathematica output of linear regression models
- Aggregated with probability and (statistical) inference



# **Probability Review:**

## **A (scalar) random variable $X$**

- Distribution: Discrete type v.s. Continuous type

More details will be covered in Probability class.



# Probability Review:

## A (scalar) random variable $X$

In this course, we mainly focus on random variables with probability density functions (pdf)

- Distribution  $\mathbb{P}[X \leq x] = \int_{-\infty}^x p(y)dy$
- Expectation  $\mathbb{E}[X] = \int_{-\infty}^{\infty} p(x)dx$
- Variance  $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 p(x)dx$





# Probability Review: Random vector (X,Y)

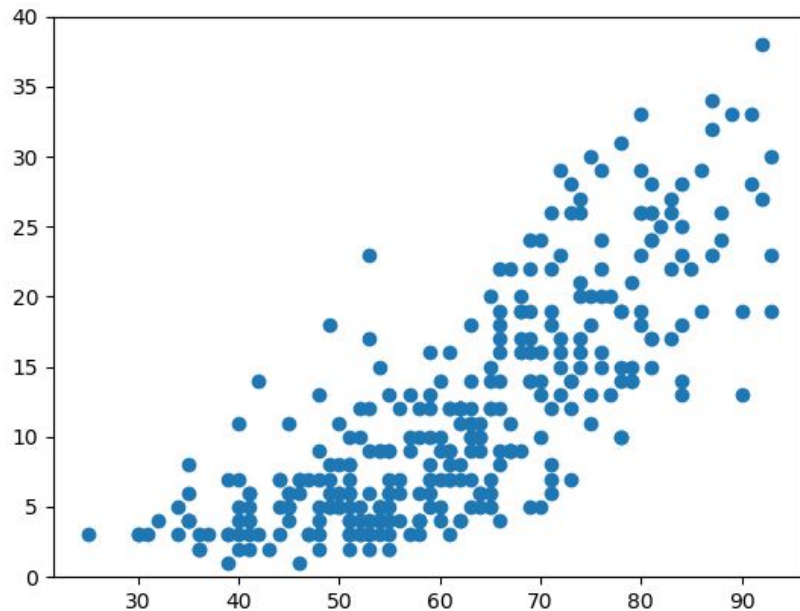
- Joint distribution  $\mathbb{P}[X \leq x, Y \leq y] = \int_{-\infty}^x \int_{-\infty}^y p_{X,Y}(x, y) dy dx$
- Marginal distribution  $p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y) dy$
- Conditional probability  $p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$
- Conditional expectation  $\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y p_{Y|X}(y|x) dy$
- Conditional variance  $Var[Y|X = x] = \int_{-\infty}^{\infty} (y - \mathbb{E}[Y|X = x])^2 p_{Y|X}(y|x) dy$

Notice that the conditional expectation and variance are both depend on the value of x. It is generally reasonable to assume that the conditional mean and variance functions are continuous.



# Regression analysis

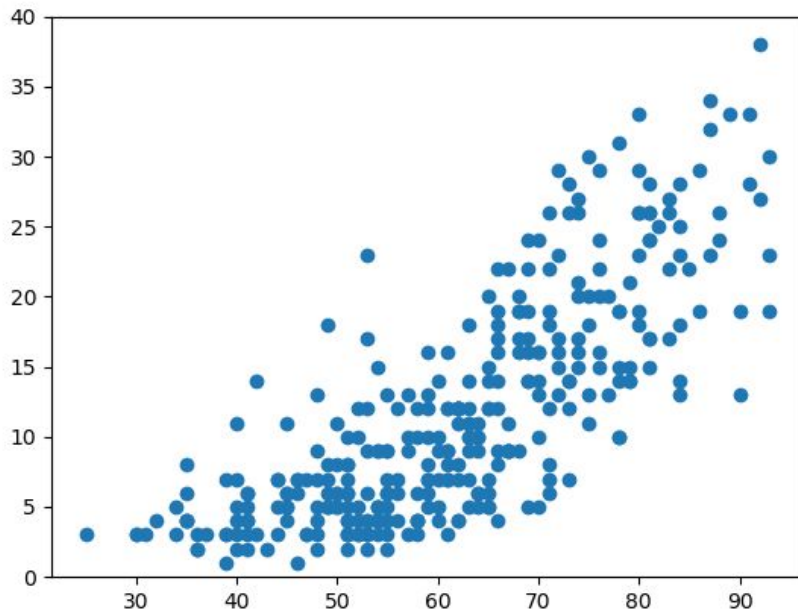
## What are we interested in?





# Regression analysis

## What are we interested in?



- We are most interested in the **conditional distribution of Y given X**.

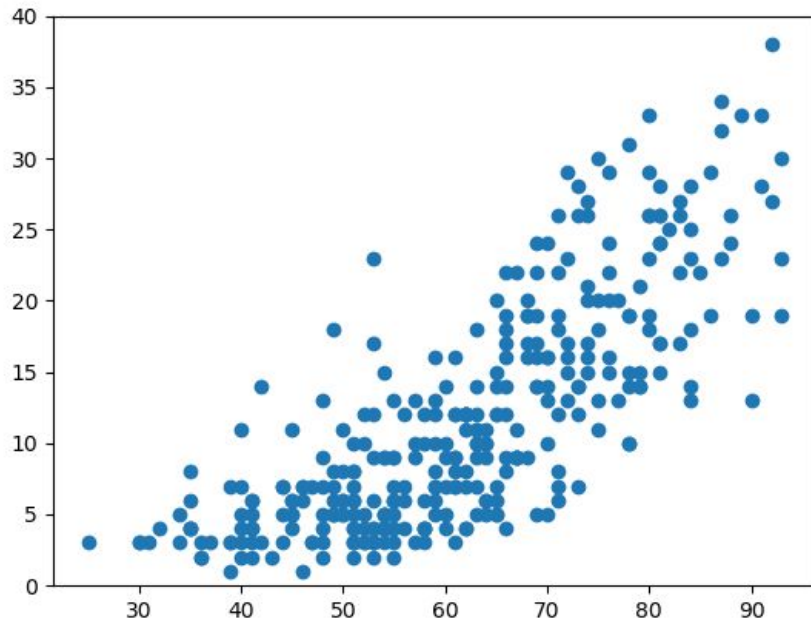
# Regression analysis

## What are we interested in?

- Typical question: what is the distribution of  $Y|X=50$  compare to  $Y|X=70$ ?
- Life gets easier if we can justify further assuming the **linear relationship**

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

$$Var[Y|X = x] = \text{constant}$$



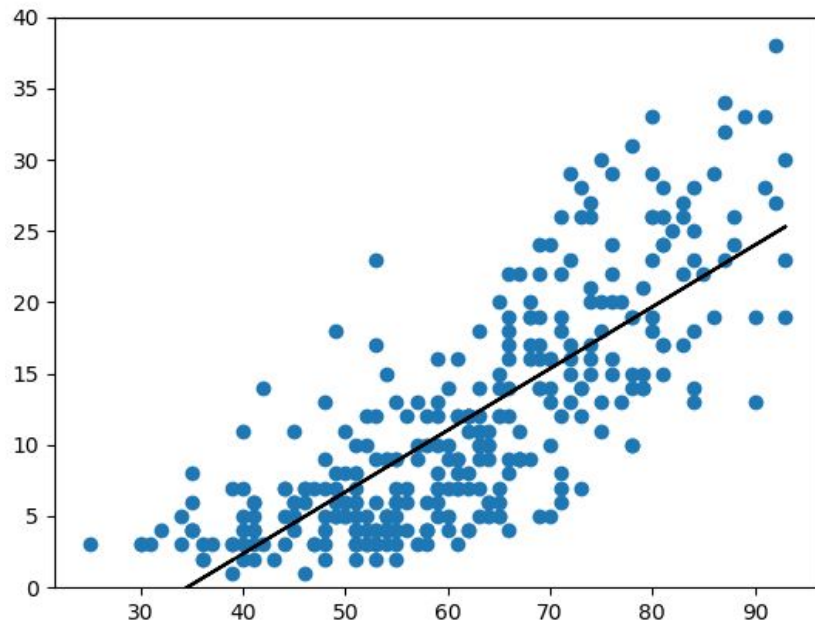
# Regression analysis

## What are we interested in?

- Typical question: what is the distribution of  $Y|X=50$  compare to  $Y|X=70$ ?
- Life gets easier if we can justify further assuming the **linear relationship**

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

$$Var[Y|X = x] = \text{constant}$$





**In other words...**

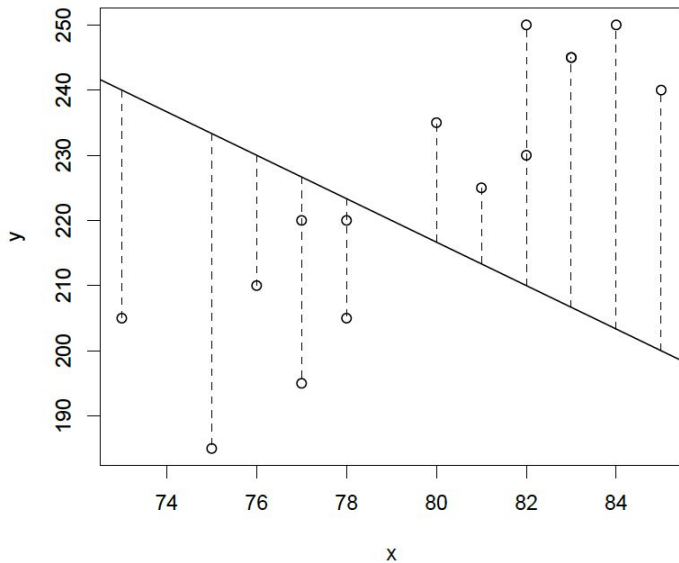
## **The Simple Linear Regression Model**

- Let  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  be samples from  $(X, Y)$
- If the SLR model holds, we write  $Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$ ,
- Here,  $\epsilon_i$  satisfies  $\mathbb{E}[\epsilon_i] = 0$  and  $\mathbb{E}[\epsilon_i\epsilon_j] = \sigma^2\delta_{ij}$
- Model parameters:  $\beta_0, \beta_1, \sigma^2$

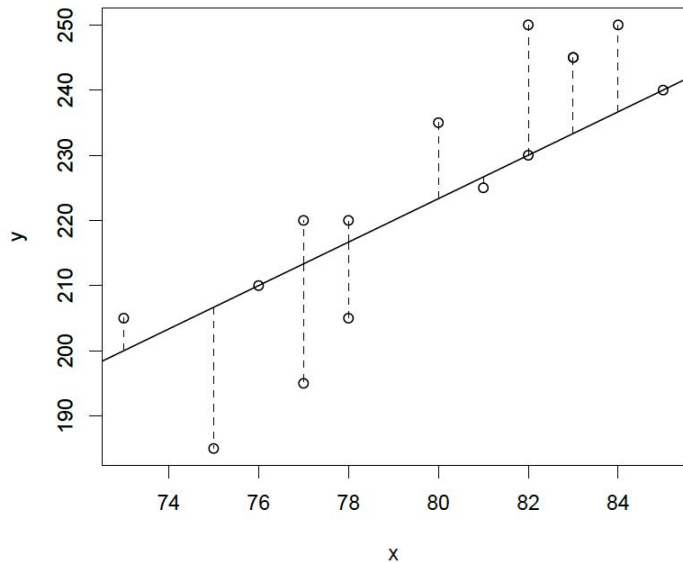


# How to fit the model?

**A bad line**



**A better line**





# Mean Squared Error (MSE)

- How to fit the model? **Mean Squared Error minimizer!**

$$\min_{\beta_0, \beta_1} \mathbb{E} [(Y - \beta_0 - \beta_1 X)^2]$$





# Mean Squared Error (MSE)

- How to fit the model? **Mean Squared Error minimizer!**

$$\min_{\beta_0, \beta_1} \mathbb{E} [(Y - \beta_0 - \beta_1 X)^2]$$

- How to express using observed data? (SLLN to approximate expectation!)

$$\min_{\beta_0, \beta_1} Q := \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$



# Mean Squared Error (MSE)

- How to fit the model? **Mean Squared Error minimizer!**

$$\min_{\beta_0, \beta_1} \mathbb{E} [(Y - \beta_0 - \beta_1 X)^2]$$

- How to express using observed data? (SLLN to approximate expectation!)

$$\min_{\beta_0, \beta_1} Q := \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

- Matrix form:

$$\min_{\beta_0, \beta_1} Q := \|y - \beta_0 - \beta_1 x\|^2$$



## Can also be found in...

- Ridge Regression

$$\min_{\beta_0, \beta_1} \|\mathbf{y} - \beta_0 - \beta_1 x\|^2 + \lambda |\beta_1|^2$$

- Lasso

$$\min_{\beta_0, \beta_1} \|\mathbf{y} - \beta_0 - \beta_1 x\|^2 + \lambda |\beta_1|$$

- Autoregression(AR) model

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n-1} (x_{i+1} - \beta_0 - \beta_1 x_i)^2$$

- And a lot of other topics in convex optimization... e.g. fused lasso/trend filtering



## References and further reading

- Kutner, Nachtsheim, Neter: *Applied Linear Regression Models* Chapter 1 & Appendix 1
- Agresti: *Foundations of Linear and Generalized Linear Models* Chapter 1