SDS7102: Linear Models and Extensions

Introduction

Qiang Sun, Ph.D. <qiang.sun@mbzuai.ac.ae>

August 19, 2025

MBZUAI

Welcome to SDS7102

- Instructors

  Qiang Sun <qiang.sun@mbzuai.ac.ae>

  Eric Moulines <eric.moulines@mbzuai.ac.ae>

  Office hours: Thursday 12:30 - 1:30 pm
- Teaching Assistant:

  Ding Bai <ding.bai@mbzuai.ac.ae>
- Course website: https://nexais.github.io/sds7102/

Basic Information

- Syllabus:
  - Available at course website/moodle.
- Textbook:
  - Lectures notes.
  - [Knight] Keith Knight. (1999) Mathematical Statistics. Chapman & Hall.
  - [PD] Peng Ding. (2025) Linear Model and Extensions. Chapman & Hall.
    - Freely available at https://arxiv.org/pdf/2401.00649v2.
- Programming language:
  - `Python`: https://www.python.org/.

Evaluation

- Evaluation
  - Lecture attendance: 10%.
  - Lab attendance: 10%.
  - Assignments: 20%.
  - Midterm Exam: 20%.
  - Final Exam: 40%.
- Homework
  - 4 assignments.
  - You are encouraged to discuss them with anyone.
  - DO NOT copy homework!
- Exams
  - 1 midterm and 1 final exam, roughly at week 8 and week 16, respectively.

Course Overview

- In the first week, we will review some basic knowledge, and introduce `Python`.
- Topics in linear models: Multivariate liear regression, statistical inference, model fitting and checking, model misspecfication, overfitting and explicit regularization, overparameterization and implicit regularization, genearalized linear models.
- Use `Python` for model fitting, simulation and numerical optimization.

Linear Models

A linear form of relationship

- In many research questions, we want to analyze the relationship between the response variable Y and some predictors $X_1, \ldots, X_p$. Examples:
  - Model height with age and gender
  - Model the risk of lung cancer with smoking status and biomarkers
- If we assume that the relationship is linear, we usually write:

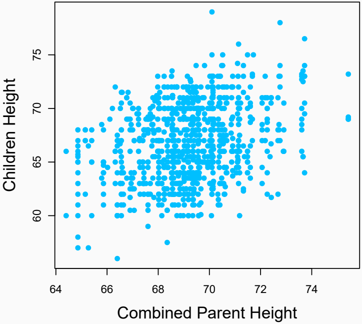
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

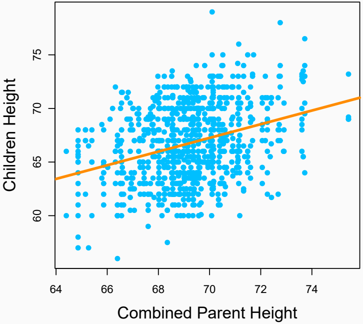
- Is this the correct relationship?

"All models are wrong, but some are useful."
— George Box (1919 – 2013)

A linear form of relationship

The heights of parents and children (Galton, 1886)





A linear form of relationship

- Why linear models?
  - Simple, can be easily interpreted
  - Approximate the truth well in practice
  - The parameters can be easily solved, and have good statistical properties
- Linear models can handle nonlinearity by incorporating nonlinear terms of covariates.
  - *Linear: Linear in parameters, not linear in covariates.*
- Linear models serve as a building block for more complicated models.

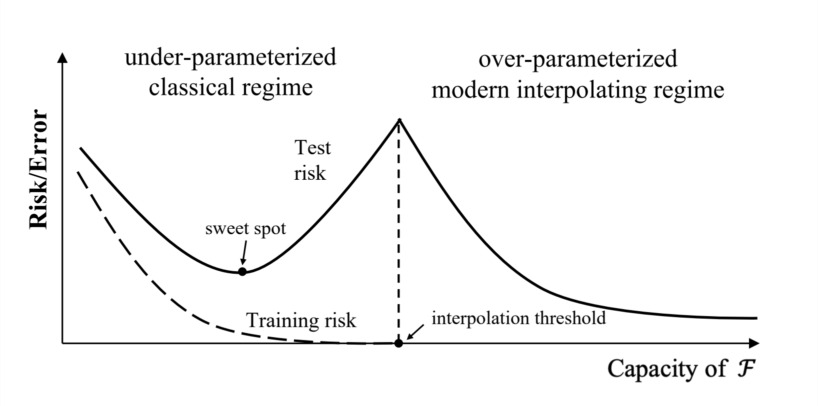
Extensions of linear models

- Y is binary, X is mixed type: logistic regression.
- Y is categorical without ordering: multinomial logistic regression (softmax head/regression).
- Y is categorial with ordering: proportional odds regression.
- Y represents counts: Poisson regression/negative-binomial regression/zero-inflated regression.
- Y is multivariate and correlated: generalized estimating equations (GEE).
- Y represents time-to-event: Cox proportional hazards regression or survival analysis more generally.
- $\ldots$

When linear models are not enough?

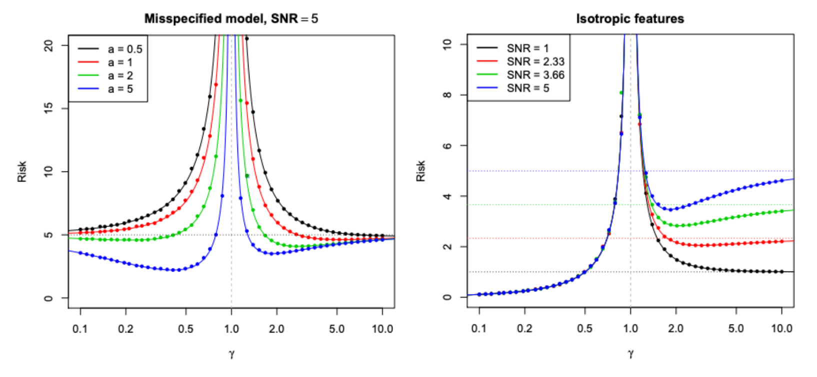
- Linear models offer insights into more complicated models, such as neural networks.
- For example, the double-descent phenomenon—originally observed empirically in deep learning (Belkin et al., 2019)—can be rigorously examined and proved within the framework of linear models (Hastie et al., 2022).

The double-descent phenomenon



Belkin et al. (2019)

The double-descent phenomenon in linear regression



Hastie et al. (2022)

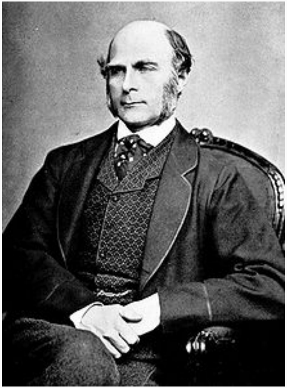
Only in rare cases,

the insights gained from linear models do not apply to more complicated models!

A brief history of linear regression

- Statistics originated from genetic studies
- Galton (Natural Inheritance, 1894) studied the diameters of mother seeds and daughter seeds, and observed a slope of 0.33 of the regression line between the two measurements (daughter seed$\sim 0.33\times$mother seed)).
- This indicates that extremely large or small mother seeds typically generated substantially less extreme daughter seeds.
- The original data can be found here.

A linear form of relationship



Francis Galton (1822 – 1911)

Karl Pearson (1857 – 1936)

A brief history of linear regression

- A formal definition of regression and correlation was developed by Karl Pearson (1896):
- The Pearson correlation is defined as:

$$\rho_{X,Y} = \text{Corr}(X,Y) = \frac{E\big[(X-E(X))(Y-E(Y))\big]}{\sigma_X \sigma_Y}.$$

- For a simple linear regression (one predictor), the slope β is

$$\beta = \text{Corr}(X,Y)\frac{\sigma_Y}{\sigma_X} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}.$$

- Multiple linear regressions are slightly more difficult.

Basic Requirements

- This course assumes basic mathematical and statistical background.
- Linear Algebra: matrix operations, linear space, operations, properties, etc.
- Calculus: double integration, etc.
- Statistical concepts: likelihood, parameter estimations, hypothesis testing, confidence intervals, central limit theorem, etc.
- Computational skills: `Python` centered, simulations, etc.

Let's try a statistical concept:

- To estimate the mean μ_x of a random variable X, one obtained a 95% confidence interval $(22.3, 25.6)$.

Let's try a statistical concept:

- To estimate the mean μ_x of a random variable X, one obtained a 95% confidence interval $(22.3, 25.6)$.
- This means (True / False ?):
  - μ_x must be between 22.3 and 25.6.

Let's try a statistical concept:

- To estimate the mean μ_x of a random variable X, one obtained a 95% confidence interval $(22.3, 25.6)$.
- This means (True / False ?):
  - μ_x must be between 22.3 and 25.6.
  - If I randomly draw another X, it has 95% chance to fall in $(22.3, 25.6)$.

Let's try a statistical concept:

- To estimate the mean μ_x of a random variable X, one obtained a 95% confidence interval $(22.3, 25.6)$.
- This means (True / False ?):
  - μ_x must be between 22.3 and 25.6.
  - If I randomly draw another X, it has 95% chance to fall in $(22.3, 25.6)$.
  - The probability that μ_x falls outside the interval $(22.3, 25.6)$ is 5%.

Let's try a statistical concept:

- To estimate the mean μ_x of a random variable X, one obtained a 95% confidence interval $(22.3, 25.6)$.
- This means (True / False ?):
  - μ_x must be between 22.3 and 25.6.
  - If I randomly draw another X, it has 95% chance to fall in $(22.3, 25.6)$.
  - The probability that μ_x falls outside the interval $(22.3, 25.6)$ is 5%.
- Where does that 95% come from?

Let's try a statistical concept:

- To estimate the mean μ_x of a random variable X, one obtained a 95% confidence interval $(22.3, 25.6)$.
- This means (True / False ?):
  - μ_x must be between 22.3 and 25.6.
  - If I randomly draw another X, it has 95% chance to fall in $(22.3, 25.6)$.
  - The probability that μ_x falls outside the interval $(22.3, 25.6)$ is 5%.
- Where does that 95% come from?
- The concept of a random variable (the CI) and its instance (the interval (22.3, 25.6)).

Some basics of the Normal distribution

Some distributions:

- Normal distribution $\mathcal{N}(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- Student's t-distribution with d.f. r:

$$f(x) = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi r}\,\Gamma(\frac{r}{2})} \left(1 + \frac{x^2}{r}\right)^{-\frac{r+1}{2}}$$

- F-distribution with d.f. d_1 and d_2:

$$f(x) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2} x\right)^{-\frac{d_1+d_2}{2}}$$

Relationships among distributions

- $X \sim \mathcal{N}(0,1) \quad \longrightarrow \quad X^2 \sim \chi^2(1)$
- $Z \sim \mathcal{N}(0,1), X \sim \chi^2(r) \quad \longrightarrow \quad \frac{Z}{\sqrt{X/r}} \sim t(r)$
- $X_1 \sim \chi^2(a), X_2 \sim \chi^2(b) \quad \longrightarrow \quad \frac{X_1/a}{X_2/b} \sim F(a,b)$

! Review the properties of Normal, χ^2, t, and F.

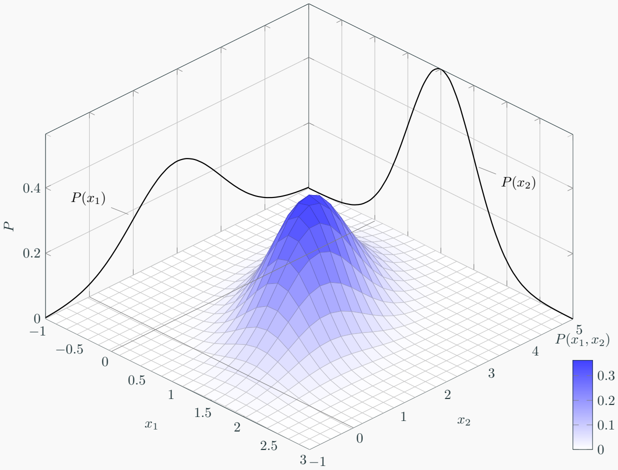
Multivariate normal distribution

- Normal (Gaussian) distribution is the most frequently used distribution in statistics
- By the central limit theory, sample means will converge to Gaussian as sample size increases
- In many cases, we will concern about two or many normally distributed random variables
- Lets consider two random variables X and Y that are jointly normally distributed with density function

$$\begin{aligned} & f(x,y) \\ = & \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \\ & \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}\end{aligned}$$

where μ_x and μ_y are the means, σ_x and σ_y are the standard deviations, and ρ is the correlation coefficient.

Multivariate normal distribution



Multivariate normal distribution

- The marginal pdfs of X and Y are also Gaussian: $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$.

? How to derive the marginal from the joint?

- What about the conditional distribution of Y given X?
- Example: Suppose that a large class took two exams. The exam scores X (Exam 1) and Y (Exam 2) follow a bivariate normal distribution with $\mu_x = 70$, $\mu_y = 60$, $\sigma_x = 10$, $\sigma_y = 15$, and $\rho = 0.6$. A student is selected at random. Suppose we know that the student got a 80 on Exam 1, what is the probability that his/her score on Exam 2 is over 75?

Multivariate normal distribution

- The question is essentially finding P$(Y > 75|X = 80)$, given that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 70 \\ 60 \end{bmatrix}, \begin{bmatrix} 10^2 & 0.6 \cdot 10 \cdot 15 \\ 0.6 \cdot 10 \cdot 15 & 15^2 \end{bmatrix}\right)$$

- We need to find the conditional density $f(y|x)$, which is defined as

$$f(y|x) = \frac{f(x, y)}{f(x)}$$

- Some derivation is required

Multivariate normal distribution

$$
\begin{aligned}
&\frac{f(x,y)}{f(x)} \\
&= \frac{\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]+\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma_y^2(1-\rho^2)}\left[\rho^2\frac{\sigma_y^2}{\sigma_x^2}(x-\mu_x)^2+(y-\mu_y)^2-2\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)(y-\mu_y)\right]\right\}}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma_y^2(1-\rho^2)}\left[y-\mu_y-\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right]^2\right\}}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}
\end{aligned}
$$

- Hence, the conditional distribution of $Y|X=x$ is

$$
\mathcal{N}\left(\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), \sigma_y^2(1-\rho^2)\right)
$$

Example:

- Hence, given that $X = 80$, the conditional distribution of Y is $\mathcal{N}(69, 12^2)$, and

$$\mathsf{P}(Y > 75 | X = 80) = \mathsf{P}(\mathcal{N}(69, 12^2) > 75) \approx 0.3085.$$

- Following the same assumption on the joint distribution of X (Exam 1) and Y (Exam 2), with $\mu_x = 70$, $\mu_y = 60$, $\sigma_x = 10$, $\sigma_y = 15$, and $\rho = 0.6$, calculate
  - Suppose we know that a randomly sampled student got 66 on Exam 1, what is the probability that the Exam 2 score is over 75?
  - Suppose we know that a randomly sampled student got 70 on Exam 2, what is the probability that the Exam 1 score is over 80?

Multivariate normal distribution

- The sum of two random normal variables are also normally distributed.
- Suppose that $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ and the correlation coefficient between X and Y is ρ, then the sum

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y)$$

- From the previous example, what is the probably that a randomly selected student has a combined score over 150, i.e., P$(X + Y > 150)$?
- Find P$(2X + 3Y > 350)$.
- Find that the student did better on Exam 1 than on Exam 2, i.e., P$(X - Y > 0)$.

Multivariate normal distribution

- We usually represent a multivariate normal (MVN) distribution in a matrix form:
- Let $X = (X_1, X_2, \dots, X_p)^\mathsf{T}$ be a p-dimensional random vector that follows the distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is symmetric and positive-definite.
- The pdf of X is

$$\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Let Z be a q-dimensional vector of linear combinations of X such that $Z = \mathbf{A}_{q\times p}X + \mathbf{b}_{q\times 1}$, then we have Z follows a MVN distribution:

$$Z \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T})$$

- A special case: if $Z = \boldsymbol{\Sigma}^{-1/2}(X - \boldsymbol{\mu})$, then entries in Z follow iid normal:

Multivariate normal distribution

- Conditional distribution of multivariate normal is also frequently used
- Let the random vector $(X^{\mathsf{T}}, Z^{\mathsf{T}})^{\mathsf{T}}$ be jointly distributed as

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{xz}^{\mathsf{T}} & \boldsymbol{\Sigma}_{zz} \end{bmatrix} \right)$$

- The the conditional distribution of $X|Z = z$ is

$$X|Z = z \ \sim \ \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xz}\boldsymbol{\Sigma}_{zz}^{-1}(z - \boldsymbol{\mu}_z), \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{xz}^{\mathsf{T}} \right)$$

Multivariate normal distribution

- Example: Suppose

$$X \sim \mathcal{N}_3 \left(\begin{bmatrix} 5 \\ 3 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & 2 \\ 0 & 2 & 9 \end{bmatrix} \right)$$

- Find P$(X_1 > 8)$
- Find P$(X_1 > 8 | X_2 = 1, X_3 = 10)$
- Find P$(4X_1 - 3X_2 + 5X_3 < 63)$
- Sometimes using `Python` to calculate these will be a lot easier.

An introduction to `Python`

Install and setup `Python`, with `VSCode`

- `Python` is a free and open-source software for computing.
- `Python` programming is usually self-explanatory, intuitive, and popular.
- `VSCode` is an integrated development environment (IDE) for `Python`.
  - `PyCharm` is another popular option.
- There are a lot of online guides available.
- We will go over some basics of the `Python` programming language.

Example

- Use `Python` on the previous example of the MVN distribution
- $\text{P}(X_1 > 8 | X_2 = 1, X_3 = 10)$

```python
import numpy as np
from scipy.stats import norm

# Define the mean vector and covariance matrix
mu = np.array([5, 3, 7])
Sigma = np.array([[4, -1, 0],
                  [-1, 4, 2],
                  [0,  2, 9]])

# Conditional mean
Mean = mu[0] + Sigma[0,1:] @ np.linalg.inv(Sigma[1:,1:]) @ (np.array([1, 10]) - mu[1:])

# Conditional variance
Var = Sigma[0,0] - Sigma[0,1:] @ np.linalg.inv(Sigma[1:,1:]) @ Sigma[1:,0]

# Compute the probability P(X1 > 8 | X2 = 1, X3 = 10)
p = norm.sf(8, loc = Mean, scale = Var)

print("P(X1 > 8 | X2 = 1, X3 = 10):", f"{p:.7f}")
```

[6] 0.0s Python

```
P(X1 > 8 | X2 = 1, X3 = 10): 0.2725755
```

Example

- P$(4X_1 - 3X_2 + 5X_3 < 63)$

```python
# mean vector and covariance matrix (same as before)
mu = np.array([5, 3, 7])
Sigma = np.array([[4, -1, 0],
                  [-1, 4, 2],
                  [0,  2, 9]])

# define the linear combination
a = np.array([4, -3, 5])

# mean and variance of a'X
Mean_aX = mu @ a
Var_aX = a @ Sigma @ a

# probability P(a'X <= 63)
p = norm.cdf(63, loc=Mean_aX, scale=np.sqrt(Var_aX))

print("P(a'X <= 63):", f"{p:.7f}")
```

[9] ✓ 0.0s Python

```
P(a'X <= 63): 0.8413447
```