Introduction to Linear Regression

The Math of Applied Statistics

- Very often, the data come in (x, y) pairs
- Given x we would like to predict y
- Many potential combinations exits...

exm: age group
0 - 3, 4 - 25, 25+

x \ y	$\mathbb{R}$	$\{0,1\}$	k categories	ordered categories	$\mathbb{R}^p$	$\mathbb{N}$	...
$\emptyset$							
$\{0,1\}$							
k categories							
ordered categories							
$\mathbb{R}$							
$\mathbb{R}^p$							
$\vdots$							

This class

Predicting from a distribution

- We want to guess (predict) the value of an unknown measurement y
- We propose a probabilistic model: the measurement is a RV $Y \sim P_Y$
- We seek to minimize

$$\mathsf{MSE}(m) := \mathbb{E}\left[(Y-m)^2\right]$$

- Set $\mu(x) := \mathbb{E}[Y]$. We have

$$\begin{aligned}
\mathsf{MSE}(m) = \mathbb{E}\left[(Y-m)^2\right] &= \mathbb{E}\left[(Y-\mu+\mu-m)^2\right] \\
&= \mathbb{E}\left[(Y-\mu)^2\right] + \mathbb{E}\left[(\mu-m)^2\right] + 2(\mu-m)\mathbb{E}[Y-\mu] \\
&= \mathbb{E}\left[(Y-\mu)^2\right] + (\mu-m)^2 + 0 \\
&= \mathrm{Var}[Y] + (\mu-m)^2
\end{aligned}$$

$(\mathbb{E}[Y]-\mu)$

$\mathsf{MSE}(m)$ is minimal when $\mu = m$.

Prediction from a conditional distribution

- Suppose a **probabilistic** model $Y \sim P_Y(x)$. The "best" predictor of y given x in the MSE sense is the **conditional expectation**:

$$\mu(x) = \mathbb{E}\left[Y|X = x\right].$$

  Indeed, using previous slide's logic:

$$\mathbb{E}\left[(Y - \mu(x))^2 \,|X = x\right] \leq \mathbb{E}\left[(Y - m(x))^2 \,|X = x\right]$$

  for any function $m(x)$
- If X is random and we have a probability model $Y, X \sim P_{X,Y}$, then

$$\mathbb{E}\left[(Y - \mu(X))^2\right] \leq \mathbb{E}\left[(Y - m(X))^2\right]$$

  The assumption $Y, X \sim P_{X,Y}$ gives rise to a **correlation model** for the dependency between the variables.

Linear Regression with One Predictor

- We restrict our prediction function $m(x)$ to have a **linear** (actually, affine) form $m(x) = \beta_0 + \beta_1 x$
- The MSE is a function of β_0 and β_1

$$\text{MSE}(\beta_0, \beta_1) = \mathbb{E}\left[(\beta_0 + \beta_1 x - Y)^2\right]$$

- We have

$\mu(x) = \mathbb{E}(Y \mid X = x)$

$$\text{MSE}(\beta_0, \beta_1) = \mathbb{E}\left[(\mu(x) - Y)^2\right] + (\mu(x) - m(x))^2,$$

so that the linear predictor is optimal iff

$$\mu(x) = \mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x,$$

In practice, this is rarely the case. George Box's dictum

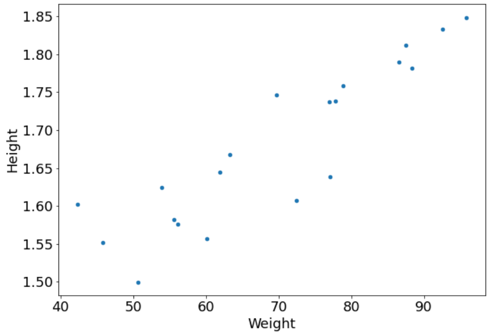
"All models are wrong, but some are useful"

comes to mind here.

Linearity

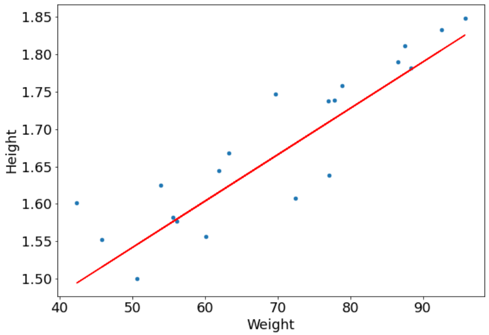
- Suppose we are given measurements of **height** and **weight** of **many** individuals

	Height	Weight
0	1.875714	109.720985
1	1.747060	73.622732
2	1.882397	96.497550
3	1.821967	99.809504
4	1.774998	93.598619
...	...	...



- We propose a model:

$$y_i = \beta_0 + \beta_1 x_i, \qquad (x_i, y_i) = (\text{height}_i, \text{height}_i)$$



Beyond Simple Linearity

- A Linear model with p **predictors** and $p+1$ **parameters**:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \ldots, n$$

  We will also use the notation

$$\mathbb{E}\left[Y|X = x\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- For example, home sale prices:

$$\begin{aligned} y_i &= \text{sale price of home } i \\ x_{i1} &= \text{square meters of home } i \\ x_{i2} &= \#\text{ of bedrooms of home } i \\ \vdots &= \vdots \\ x_{i,203} &= \#\text{ of synagogues near home } i \end{aligned}$$

- Remarks:
  - The model is **linear** in $\beta = (\beta_0, \ldots, \beta_p)$, not in x
  - Would **still be linear** if we add $x_{i,204} = \sqrt{\#\text{of bedrooms}}$
  - **Sum** of linear models is also a **linear model**

Advanced Stats. for DS 1/3/2022

Lecture 1

We started with the following slides:

- The math of applied stat.
- Predicting from a distribution
- Predicting one RV from another
- Linear regression with One Predictor

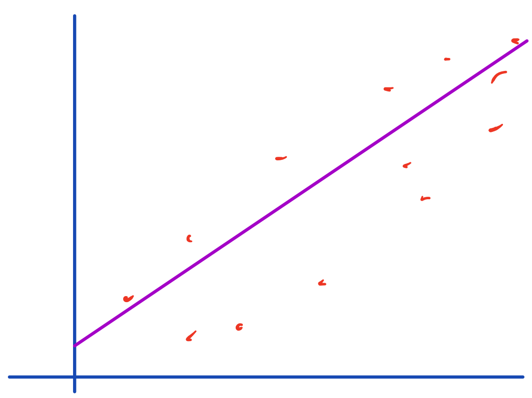
Linearity

suppose you are given measurements of height weight of many individuals

id	Height [cm]	Weight [kg]
1	180	109.7
2	174	73.6
⋮	⋮	⋮

We propose a model:

$y_i = \beta_0 + \beta_1 x_i$



$y_i = \text{weight}_i$

$x_i = \text{height}_i$

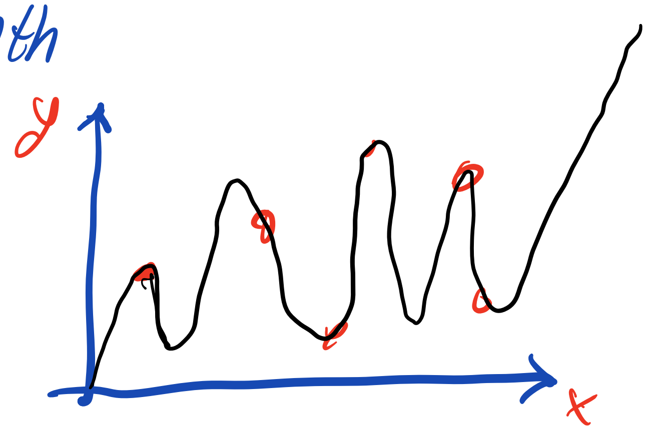
Polynomial Regression

$$y_i = \beta_0 + \beta_i x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k + \varepsilon_i \quad x_i \in R$$

in short:

$$E[Y|X=x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k \quad x \in \mathbb{R}$$

- makes sense if the relationship between x and y is smooth



- Given data, we can approximate it arbitrarily well for large k (zero error if $k=n-1$)
- Perfect apfx in linear models is suspicious, usually indicates an overfit.

Two Groups

- suppose we want to compare two groups: male/female, nickel vs copper, treatment vs control

- We encode one of the group as 0 and the other one as 1:

for example:

$$(*) \quad E[Y|X=x] = \begin{cases} \beta_0 + \underbrace{\beta_1}_{\text{extra effect}} & x=1 \\ \beta_0 & x=0 \end{cases}$$

- We can write (*) as

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

Notation: dummy variable

---

k groups

$$x_1 = \begin{cases} 1 & \text{if group 1} \\ 0 & \text{otherwise} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if group 2} \\ 0 & \text{otherwise} \end{cases} \quad \ldots, x_{k-1} = \begin{cases} 1 & \text{if group } k-1 \\ 0 & \text{otherwise} \end{cases}$$

- We get:

$$E[Y|X=x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1}$$

(group 0 has mean β_0, mean of group $j>0$ is $\beta_0 + \beta_j$)

- Equivalently:

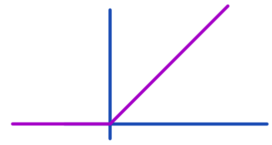
$$E[Y|X=x] = \beta_0 + \beta_1 1_{\{x=1\}} + \beta_2 1_{\{x=2\}} + \ldots + \beta_{k-1} 1_{\{x=k-1\}}$$

Two-Phase Regression

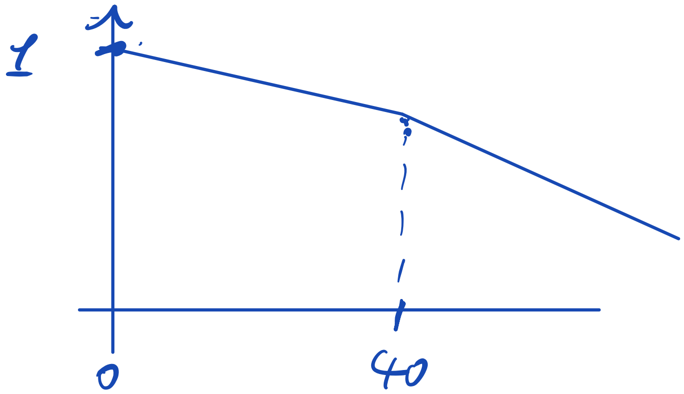
- The slope of the line changes at a certain point x_0. For example, the performance

of an average human kidney is starting to decline at age 40.
We express this as follows.

$$E[Y|X=x] = \beta_0 + \beta_1 x + \beta_2 [x - \overset{\text{fixed point}}{x_0}]_+$$



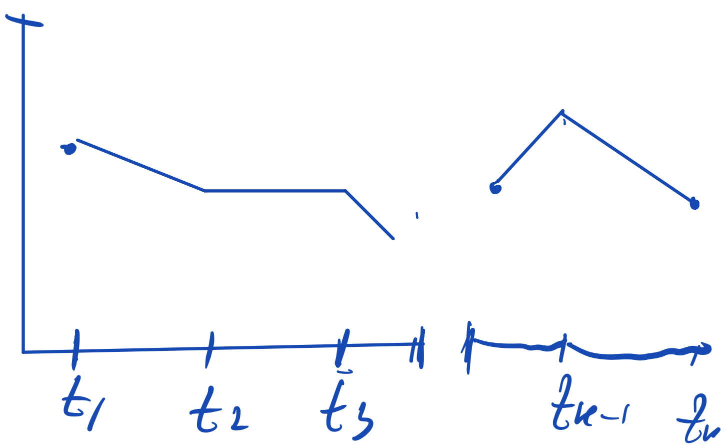
$$z_+ := \max\{0, z\} = z \cdot 1_{z>0}$$



Multiple Regression

- Suppose that we want a relationship that changes over time; time goes for k periods.
  We can use:

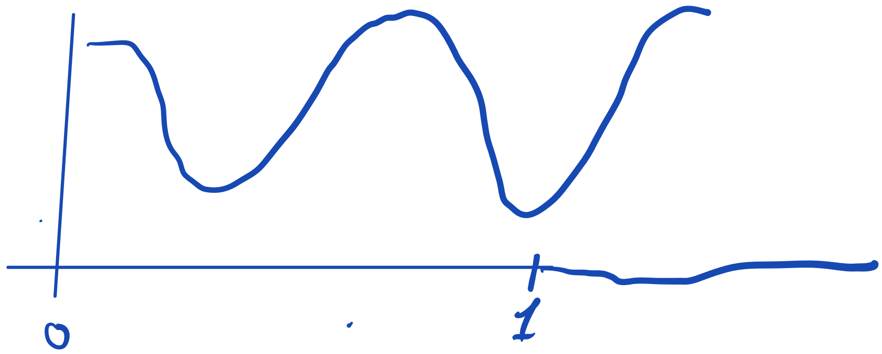
$$E[Y|X=x] = \beta_0 + \beta_1 (x - t_1)_+ + \beta_2 (x - t_2)_+ + \dots + \beta_k (x - t_k)_+$$



Periodic Functions

How can we handle cyclical data, e.g. calender time?

$$E[Y|X=x] = \beta_0 + \beta_1 \sin(2\pi f_0 x) + \beta_2 \cos(2\pi f_0 x) + \beta_3 \sin(2 \cdot 2\pi f_0 x) + \dots$$



Example: we want to predict traffic at a specific hour of the day based on features: time of day, day of week,

$$E[Y|X=x] = \beta_0 + \beta_1 \sin\left(2\pi \frac{x}{24}\right) + \beta_2 \cos\left(2\pi \frac{x}{24}\right) + \beta_3 \sin\left(2\pi \cdot \frac{x}{7 \cdot 24}\right) + \beta_4 \cos\left(2\pi \frac{x}{7 \cdot 24}\right)$$

Concluding Remarks

- despite the models' differences,

the underlying math is all linear

- Examples of non-linear models:
  - $\mathbb{E}[Y|X=x] = \beta_0(1 - e^{-\beta_1 x})$
  - $\mathbb{E}[Y|X=x] = \beta_1 x_1 + \beta_2 (x_2 - \beta_3)_+$
  - $\mathbb{E}[Y|X=x] = \sum_{j=1}^{k} \beta_j e^{-\frac{1}{2}\|x-\mu_j\|^2}$
  - $\mathbb{E}[Y|X=x] = \beta_0 + \beta_1 \sin(2\pi(x-\beta_2))$