Week 2: Linear Regression & Some Fundamental Notions and Tools in Machine Learning

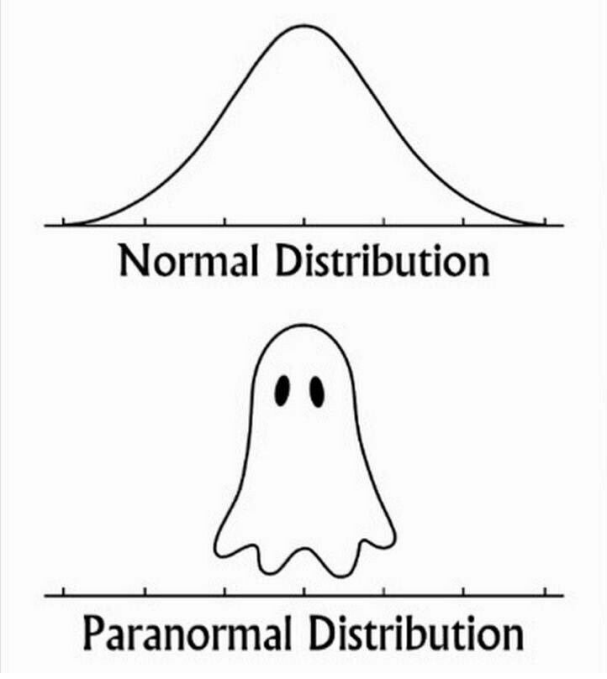
AI 539: Machine Learning for Non-Majors

Alireza Aghasi

Oregon State University

This Lecture ...

– Would be surprised to know how much of Statistics is about Normality!



Some Basic Probability Overview

- For a continuous random variable X we often define a probability density distribution $f_X(x)$ where $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$.
- A random variable X is normally distributed with mean μ and variance σ^2 (denoted as $\mathcal{N}(\mu, \sigma^2)$), when

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

- Sum of normals: if $x_1, \ldots, x_n$ are normal (**not necessarily independent**) then the weighted sum $\alpha_1 x_1 + \ldots \alpha_n x_n$ is also normal
- Expectation of the weighted sum: if $x_1, \ldots, x_n$ are random variables with mean $\mathbb{E}(x_i) = \mu_i$, then for constants α_i:

$$\mathbb{E}(\alpha_1 x_1 + \ldots \alpha_n x_n) = \alpha_1\mu_1 + \ldots + \alpha_n\mu_n$$

- Variance of the weighted sum: if $x_1, \ldots, x_n$ are **independent** random variables with variance $var(x_i) = \sigma_i^2$, then

$$var(\alpha_1 x_1 + \ldots \alpha_n x_n) = \alpha_1^2\sigma_1^2 + \ldots + \alpha_n^2\sigma_n^2$$

$X_1 : \mathbb{E}X_1 = 1 \quad , \quad \mathrm{Var}(X_1) = 2$

$X_2 : \mathbb{E}X_2 = 2, \quad \mathrm{Var}(X_2) = 4$

$$X = 3 - X_1 + 2X_2$$

$$\mathbb{E}[X] = 3 - 1 + 2 \times 2 = 6$$

$$\mathrm{Var}(X) = 0 + (-1)^2 \times 2 + 2^2 \times 4 = 18$$

A Brief Review of Hypothesis Testing

Hypothesis Testing

- A hypothesis is often a conjecture about one or more populations
- To prove a hypothesis is true we need to examine all the population which is often not practical, **instead we take a random sample** and probabilistically assess if we have enough evidence to support a hypothesis
- **Example**: All human beings respond well to a specific treatment H_0 ?
  - Instead of testing it on all people on the planet, we look into a fraction of people who are infected by the decease and see if the treatment works
- Since we are focused on a limited sample, we can only state our confidence about the conjecture probabilistically

H_1: not everyone respond well

Hypothesis Testing

- Hypothesis testing is often formulated in terms of two hypotheses
  - H_0: the null hypothesis
  - H_1: the alternate hypothesis
- The hypothesis we want to test is if H_1 is likely to be true
- Often, the equality hypothesis is chosen to be the null hypothesis
- For many problems that we encounter in this course, Hypothesis Testing is simply testing the chances of a random variable to be in regions defined by H_0 or H_1 (wait for the numerical example)

Hypothesis Testing

α { is the confidence of the test
smaller is better

① assert H_0

② $1-\alpha \geq 95\%$?

- We want to test if H_1 is likely to be true
- You decide to assert something about the null hypothesis H_0 and be certain what you assert is true with probability at least $1 - \alpha$
- **Example**: We are 90% confident that this drug works on patients with xxx decease
- So normally you decide on how confident you want to make an assertion and determine a value for α

control the confidence

Hypothesis Testing

In hypothesis testing only one of these two cases happens:

1. – You **reject** H_0 and **accept** H_1 since you have enough evidence in favor of H_1
2. – You **fail to reject** H_0, since you don not have enough evidence to support H_1

While you see a lot of documents talking about "*accepting* H_0", technically you should use the term "*fail to reject*"

- It might be the case that H_0 is false, but your data is not enough to reject it (does not mean you should accept it)

Example: H_0: Tim is innocent H_1: Tim is guilty

If you have enough to support Tim is guilty, you reject H_0. If you do not have enough to show that Tim is guilty (failure to reject H_0), that does not mean he is innocent (accepting H_0)

Hypothesis Testing: Types of Error

	Reject H_0 (accept H_1)	Do not reject H_0
H_0 true in reality (probability)	Type I error α	Correct decision $1-\alpha$ $\geq 95\%$
H_1 true in reality (probability)	Correct decision $1-\beta$	Type II error β

- α is a small number that we determine and is called the significance level (the probability of making type I error)
- We decide on how confident we want to make a claim in favor of H_0 and $1-\alpha$ is our confidence about this
- Normally people take α to be 0.05 or 0.01, giving you 95% or 99% chance of validity in making the argument in support of H_0
- We also have a type II error (calling $1-\beta$ the power of the test), but here we do not want to focus on that

Hypothesis Testing Example

- In the context of our linear regression problem we are interested in hypothesis testing problems on the basis of samples

Example: There is a normal distribution with variance 1 and unknown mean μ. We take 10 independent samples x_i of this distribution as:

`1.8978, 1.7586, 2.3192, 2.3129, 1.1351, 1.9699, 1.8351, 2.6277, 3.0933 , 3.1093`

we add up these numbers and divide it by 10 (taking the sample mean) and observe that

$$\frac{x_1 + \ldots x_{10}}{10} = 2.2059.$$ take average

We get a feeling that probably $\mu = 2$, so we decide to test this hypothesis:

$$H_0 : \mu = 2 \qquad vs \qquad H_1 : \mu \neq 2 \text{ (two sided test).}$$

assertion counter assertion

Hypothesis Testing Example

Solution: We generally look into the behavior of the random variable

$$\bar{x} = \frac{x_1 + \ldots + x_{10}}{10}$$

$\bar{X} \sim N(\mu, 0.1)$

which is a normally distributed random variable with mean μ and variance 0.1 [can you say why?]. As a result, $z = \frac{\bar{x}-\mu}{\sqrt{0.1}}$ is a standard $\mathcal{N}(0,1)$ random variable [can you say why?].

We refer to z as the **test statistic**.

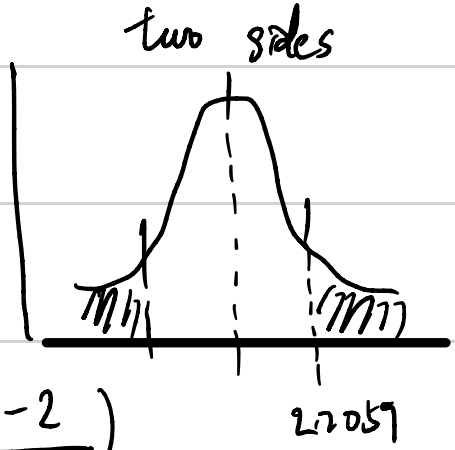
p-value: *is a useful quantity in the analysis of the test and is the probability of obtaining a result equal or "more extreme" than what we have observed, given that the null hypothesis is true. In the case of this example:* **assuming that $\mu = 2$, we perfectly know the distribution of $\bar{x}$ (or z) and want to asses how likely it is to draw a sample this far towards the tails and further (i.e., away from the mean of $\bar{x}$)**:

$$\begin{aligned}\text{p-value} =& \mathbb{P}\left(\bar{x} \geq 2.2059, \bar{x} \leq 2\mu - 2.2059 \mid \mu = 2\right) \\ =& \mathbb{P}\left(|z| \geq \frac{2.2059 - \mu}{\sqrt{0.1}} \,\middle|\, \mu = 2\right) = \mathbb{P}\left(|z| \geq 0.6511\right) = 0.5150.\end{aligned}$$

$\bar{X} \sim N$

$$E(\bar{X}) = E\left[\frac{1}{10}X_1 + \cdots + \frac{1}{10}X_{10}\right] = \frac{1}{10}E[X_1] + \cdots + \frac{1}{10}E[X_{10}] = \frac{1}{10}\mu + \cdots + \frac{1}{10}\mu = \mu$$

$$Var[\bar{X}] = Var\left[\frac{1}{10}X_1 + \cdots + \frac{1}{10}X_{10}\right] = \frac{1}{100}$$



P-value:

$$= P(\bar{X} \geq 2.2059,\ \bar{X} \leq 2\mu - 2.2059 \mid \mu = 2)$$

$$= P\left(\frac{\bar{X}-2}{\sqrt{0.1}} \geq \frac{2.2059-2}{\sqrt{0.1}},\ \frac{\bar{X}-2}{\sqrt{0.1}} \leq \frac{4-2.2059-2}{\sqrt{0.1}}\right)$$

$$Z = \frac{\bar{X}-2}{\sqrt{0.1}} \Rightarrow P\left(|Z| \geq \frac{2.2059-2}{\sqrt{0.1}}\right) = 0.515$$

Hypothesis Testing Example

Suppose our significance level is $\alpha = 0.05$.

> If p-value$\leq \alpha$: reject H_0 (accept H_1)
> If p-value$> \alpha$: fail to reject H_0

For our example p-value $= 0.5150 > 0.05$, so we cannot reject the hypothesis that $\mu = 2$. Not enough prove for $\mu = 2$

- If the value of $\bar{x} = 2.2059$ was obtained by taking the sample mean over 100 samples then we had $z = \frac{\bar{x}-\mu}{\sqrt{0.01}}$ and

$$\text{p-value} = \mathbb{P}\left(|z| \geq \frac{2.2059 - \mu}{\sqrt{0.01}} \,\middle|\, \mu = 2\right) = \mathbb{P}\left(|z| \geq 2.059\right)$$
$$= 0.0395 < 0.05,$$

  then we were able to reject H_0.
- In other words we are more than 95% confident (accurately, 96.05% confident) that it is not possible to take the sample mean over 100 random numbers of mean $\mu = 2$ and variance 1, and get a value as far from 2 as 2.2059! [Lets try it on Matlab]

Hypothesis Testing Example: Unknown Variance

Suppose in the previous example we did not know σ and instead of working with the standard random normal variable $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ we work with the random variable

$$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \qquad \text{where :} \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

for t-distribution

Note that s is an unbiased estimate of the standard deviation (since we don't know σ this is the best we can use).

If we want to go through a similar hypothesis test, we need to look into the test statistic $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ which is no more normally distributed. It has a more complicated distribution called *Student's t distribution*.

So for the probabilities needed to calculate the p-value, we need to refer to t-distribution tables instead of the normal distribution tables.

Your Take Away from Hypothesis Testing

- We have some independent samples from a distribution and we guess something about our data
- We form a hypothesis test with some null and alternate hypotheses
- We fix some value for the significance α, meaning that $1 - \alpha$ is how confident we want to be in making our claim
- We form a test statistic (the resulting random variable can have a very complicated distribution)
- We calculate the p-value:
  - If p-value$\leq \alpha$: reject H_0 (accept H_1)
  - If p-value$> \alpha$: fail to reject H_0

Now Lets Start Linear Regression!

Introduction to Linear Regression

- You remember we had an ideal "regression function" $f(\boldsymbol{x})$, which was the actual function behind our data generation and our observations y were in the form

$$y = f(\boldsymbol{x}) + \epsilon$$

- Note that our input vector $\boldsymbol{x}$ contained all the input features, i.e.,

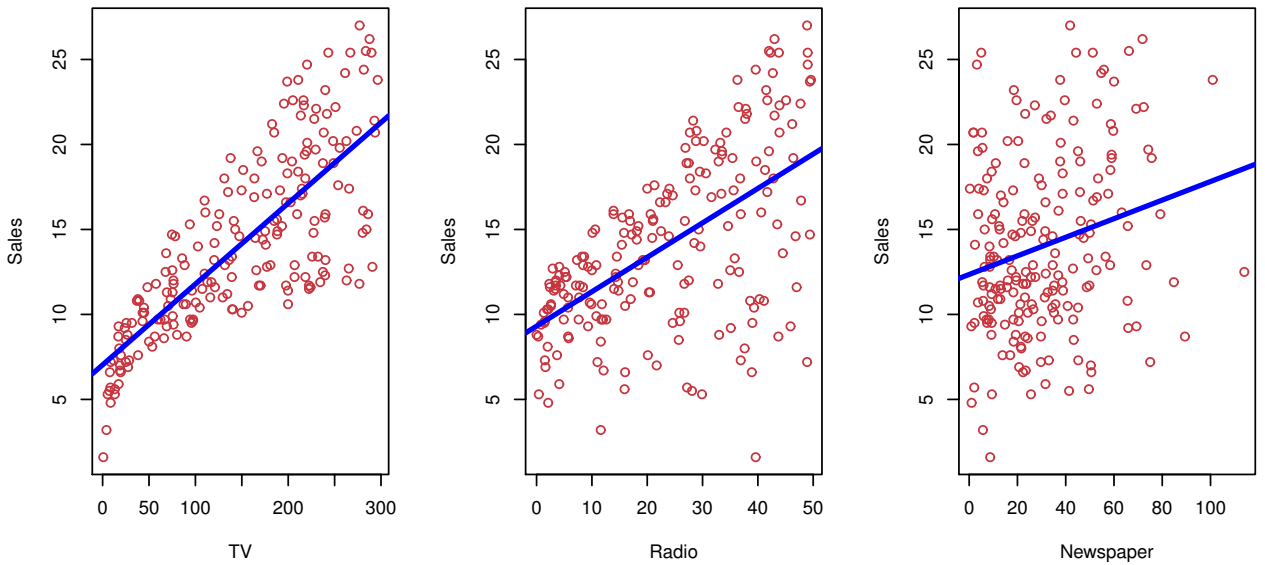
$$\boldsymbol{x} = [x_1, \ldots, x_p]^\top$$

- We had no access to $f(\boldsymbol{x})$, but we wanted to estimate it with some function $\hat{f}$
- In **linear regression** we search for a $\hat{f}$, which takes the following form

$$\hat{f}(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Introduction to Linear Regression

Lets start with a very simple model that only uses one feature



$$\text{Sales} = \beta_0 + \beta_1 \text{TV} + \epsilon$$

one factor

one time

or more generally,

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where x is the only available feature

More on Simple Linear Regression

- We consider the model

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

- β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients and ϵ is the error term.
- We are given samples of the form $(x_1, y_1), \ldots, (x_n, y_n)$, using which we try to fit some values $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients
- After this fit we can predict future responses to a test sample x_t, using

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t,$$

Determining the Model Coefficients

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$$

– Remember that we had samples $(x_1, y_1), \ldots (x_n, y_n)$ and we would like to determine $\hat{f}(x)$ by

$$\min \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

– Using our simple model we would like to decide β_0 and β_1 such that the **Residual Sum of Squares** (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

is minimized

find β_0, β_1 to minimize $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = L(\beta_0, \beta_1)$

$$0 = \frac{\partial L}{\partial \beta_0} = 2 \cdot \sum_{i=1}^{n} (-y_i + \beta_0 + \beta_1 x_i) = \sum_i \beta_0 + \beta_1 \underbrace{\sum_i x_i}_{n \cdot \bar{x}} = \underbrace{\sum y_i}_{n \cdot \bar{y}}$$

$$0 = \frac{\partial L}{\partial \beta_1} = 2 \sum_{i=1}^{n} x_i (\beta_0 + \beta_1 x_i - y_i) = \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 = \sum x_i y_i$$

$$\Rightarrow \begin{cases} n \cdot \beta_0 + \left(\sum_i x_i\right) \beta_1 = \sum_i y_i \\ \left(\sum_i x_i\right) \beta_0 + \left(\sum_i x_i^2\right) \cdot \beta_1 = \sum x_i y_i \end{cases}$$

Determining the Model Coefficients

– Taking the derivative with respect to β_0 and β_1 and setting it to zero, for $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ we get

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

– We may use the simple equalities:

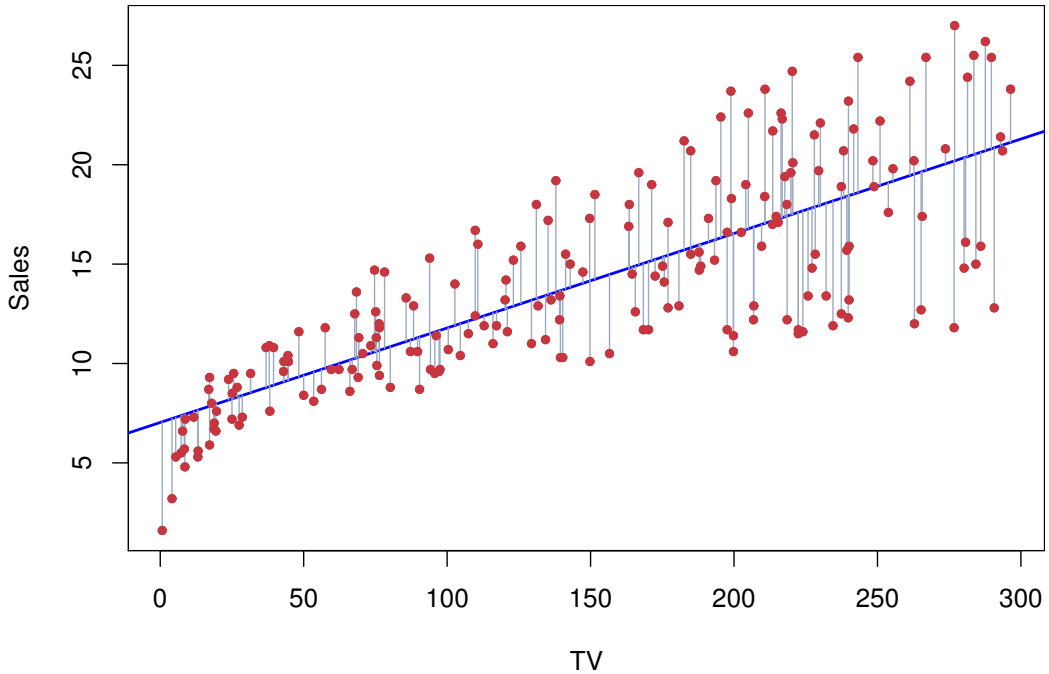
$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

to get the final equations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

just for one feature ?

How Well Did We Do the Fit?



Now that we have our fit we would like to address few questions about it!

(`We code up the answer to each question in the Example Code 1`)

Baseline: $y = \beta_0 + \beta_1 x$

Gold's Model

The estimates are unbiased: $\begin{cases} E[\hat{\beta}_1] = \beta_1 \\ E[\hat{\beta}_0] = \beta_0 \end{cases}$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Given $Var(\varepsilon) = \sigma^2 \rightarrow Var(y_i) = \sigma^2$

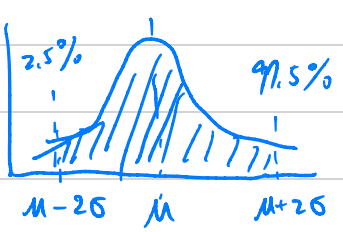
$$Var(\hat{\beta}_1) = Var\left[\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \right]$$

$$= Var\left[\frac{(x_1 - \bar{x}) y_1}{\sum [x_i - \bar{x}]^2} + \cdots + \frac{(x_n - \bar{x}) y_n}{\sum [x_i - \bar{x}]^2} \right]$$

Since all y_i are independent to each other

$$= \sigma^2 \frac{(x_1 - \bar{x})^2}{\left[\sum (x_i - \bar{x})^2\right]^2} + \cdots + \sigma^2 \frac{(x_n - \bar{x})^2}{\left[\sum (x_i - \bar{x})^2\right]^2}$$

$$= \sigma^2 \frac{\sum (x_i - \bar{x})^2}{\left[\sum (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$



What is the Confidence Interval for the Coefficients Obtained?

- Notice that $\hat{\beta}_1$ and $\hat{\beta}_0$ are both normally distributed when the noise is normally distributed
- We can define confidence intervals that the true β_1 and β_0 are in it with 95% confidence
- For $\sigma^2 = \text{var}(\epsilon)$, we define

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right), \quad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Hint on derivation: since $\sum_i (x_i - \bar{x})\bar{y} = 0$, we can start with the alternative formulation $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$ and use the properties stated at the beginning of slides to derive $var(\hat{\beta}_1)$
- The 95% confidence intervals for β_0 and β_1 are

$$\left[\beta_0 - 2\text{SE}(\hat{\beta}_0), \beta_0 + 2\text{SE}(\hat{\beta}_0) \right], \quad \left[\beta_1 - 2\text{SE}(\hat{\beta}_1), \beta_1 + 2\text{SE}(\hat{\beta}_1) \right]$$

- Exercise: Show $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased.

$$E[\hat{\beta}_1] = E\left[\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}\right] = E\left[\frac{\sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum (x_i - \bar{x})^2}\right]$$

$$= \frac{\beta_0 \cdot \sum (x_i - \bar{x}) + \beta_1 \sum (x_i - \bar{x}) x_i + 0}{\sum (x_i - \bar{x})^2}$$

$$= \beta_1 \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} = \beta_1$$

Is There a Relationship Between x and y?

- We would like to know if there is really a relationship between x and y or if the fit is useless?
- We form a hypothesis testing for β_1 (if it is zero, then x and y are not related):
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$ → There are some relationships
- Our test statistic is chosen to be $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$
- We look this up in the t-distribution table and find the p-value

(see the code)

If p-value$\leq \alpha$: reject H_0 (accept H_1)
If p-value$> \alpha$: fail to reject H_0

- For the example provided p-value $= 2 \times 10^{-16}$ and we reject H_0 (meaning that x and y are related)

How Well does the Model Explain the Data?

- We can also answer the question of how well our fitted model explains the data by defining another statistic:

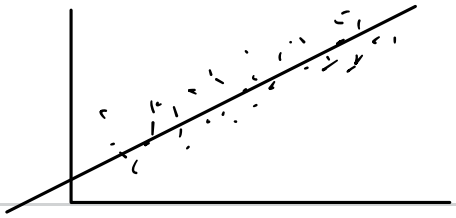
$$R^2 = 1 - \frac{RSS}{TSS}$$

where TSS is the Total Sum of Squares

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2, \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

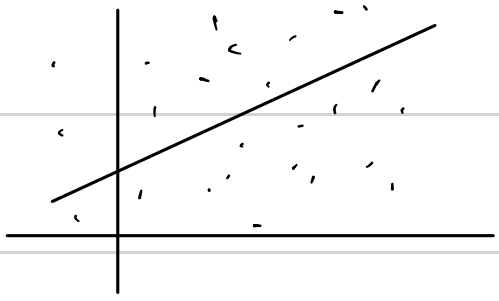
- For general regression problems (not only the simple one) R^2 statistic measures the proportion of variability in y that can be explained by x
- R^2 close to 1 indicates that our model explains a large proportion of the response variability, and R^2 close to zero indicates that our model cannot explain much of the variability in response

(see the code)



$R^2 = 1$

P-value small is good.



Multiple Linear Regression

Multiple Linear Regression

– It can be the case that we have multiple features $x_1, \ldots, x_p$ and we would like a fit like

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \epsilon$$

For example: $\text{sales} = \beta_0 + \beta_1.\text{TV} + \beta_2.\text{radio} + \beta_3.\text{newspaper} + \epsilon$

– Suppose that we have n training/response samples $(\boldsymbol{x}^{(1)}, y_1), \ldots, (\boldsymbol{x}^{(n)}, y_n)$ where

$$\boldsymbol{x}^{(1)} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_p^{(1)} \end{pmatrix}, \quad \boldsymbol{x}^{(2)} = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ \vdots \\ x_p^{(2)} \end{pmatrix}, \quad \ldots \quad , \boldsymbol{x}^{(n)} = \begin{pmatrix} x_1^{(n)} \\ x_2^{(n)} \\ \vdots \\ x_p^{(n)} \end{pmatrix}$$

Multiple Linear Regression

– We would like to minimize the following squared error

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_1^{(i)} - \beta_2 x_2^{(i)} \ldots - \beta_p x_p^{(i)} \right)^2$$

– Consider using the following matrices/vectors

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{X}_{n \times (p+1)} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{pmatrix}$$

note that matrix $\boldsymbol{X}$ has the samples along the rows

– Then, it is straightforward to see that

$$RSS = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \| y - X\beta \|^2$$

$$RSS = (y - X\beta)^T (y - X\beta)$$

$$= (y^T - \beta^T X^T)(y - X\beta)$$

$$= y^T y - 2y^T X\beta + \beta^T X^T X\beta$$

$$\frac{\partial RSS}{\partial \beta} = 0 - 2X^T y + 2X^T X\beta = 0$$

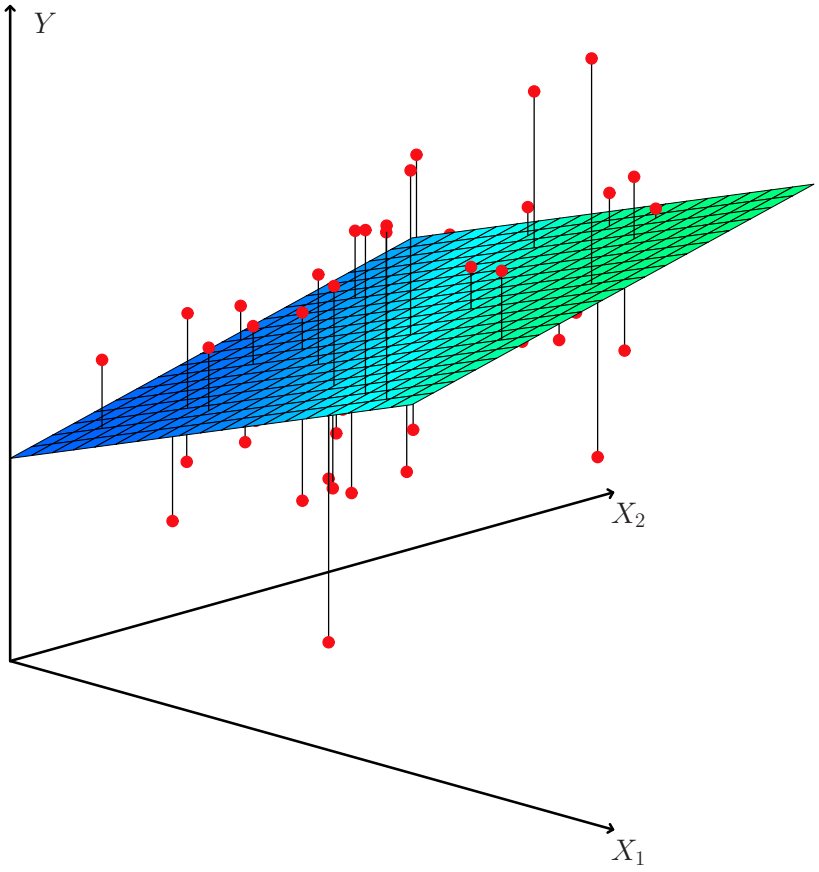
$$\Rightarrow X^T X\beta = X^T y \Rightarrow \beta = (X^T X)^{-1} X^T y$$

Multiple Linear Regression

– Similar to what we did before we can set $\partial RSS/\partial \boldsymbol{\beta} = 0$ (requires little bit of knowing how to do vector/matrix derivatives) and get

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

minimize $\|y - X\beta\|^2$
β



Are the Features and Response Related?

- We would like to know if at least one of the features $x_1, \ldots, x_p$ is useful in predicting the response.
- We form a hypothesis testing as follows:
  - $H_0 : \beta_1 = \beta_2 = \ldots = 0$
  - H_1 : at least one β_i is non-zero
- For $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$, our test statistic is chosen to be

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

which turns to have an F distribution

If p-value$\leq \alpha$: reject H_0, If p-value$> \alpha$: fail to reject H_0

- We can either use F-distribution tables and find the p-value; or use this rule: *if F is much larger than 1, we reject H_0; if F is very close to 1, we fail to reject H_0*

(see the code)

F much $> 1 \Rightarrow$ P-value $\downarrow$: Good

F close to $1 \Rightarrow$ P-value $\uparrow$: Bad

Now that we have our fit, again we would like to address few questions about it!

(We code up the answer to each question in the Example Code 2)

Intercept : β_0

If I drop either "indus" or "age"

It's highly possible to improve the other p-value?

Assessing the P-Values and Correlations Among Features

– Sometimes (most of the times) features are correlated and the contribution of one feature can be taken care of by the others

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

should be dropped

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

(see the code)

What are the Best Selection of Features?

- As noted sometimes some features can be redundant and we would like to find the best subset of features that predicts well and is not redundant
- In general this problem is "NP-hard" (computationally very hard) and we need to assess 2^p models
- There are some heuristics to do this that we will see later:
  - Forward selection: model p regressions each with only one feature, pick the one with least RSS, repeat it with selected feature and combination of others, ...
  - Backward selection: Start with all features and remove variable with largest p-value, run a new regression, remove variable of largest p-value, ...

In general, Backward is better? less time?

How Can We Handle Categorical Features?

- Sometimes our features do not take numerical values, instead they take categorical values
- **Example**: In a regression problem we have a feature called `ethnicity`, which takes possible values of `Asian, Caucasian, African-American`
- We can introduce 2 dummy variables (features) e_A, e_C
  - $e_A = 1,\ e_C = 0$ if Asian
  - $e_A = 0,\ e_C = 1$ if Caucasian
  - $e_A = 0,\ e_C = 0$ if African-American
- Basically, for every categorical feature that has L levels, we need to define $L - 1$ dummy variables

Can We Only Fit Linear Curves with Linear Regression?

Yes, we can!

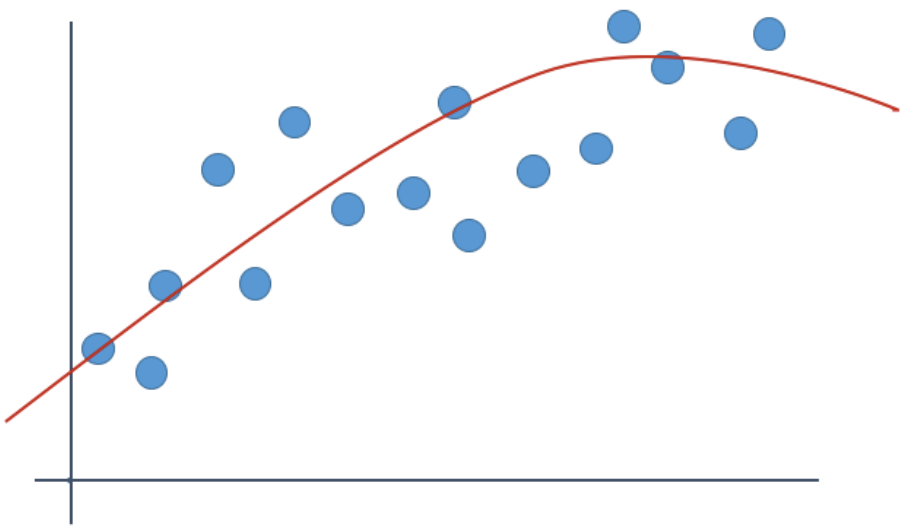
- Based on the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \epsilon$$

  one might get the impression that linear regression is only good for fitting flat surfaces (linear manifolds)
- If we include powers of a feature, e.g., $x_1, x_1^2, \ldots$ or cross terms between the features, e.g., $x_1 x_2, x_2 x_3 x_5$, etc, then we can also fit nonlinear surfaces
- Of course knowing what powers or what cross terms to include in the feature list is not always clear

Can We Only Fit Flat Curves with Linear Regression?

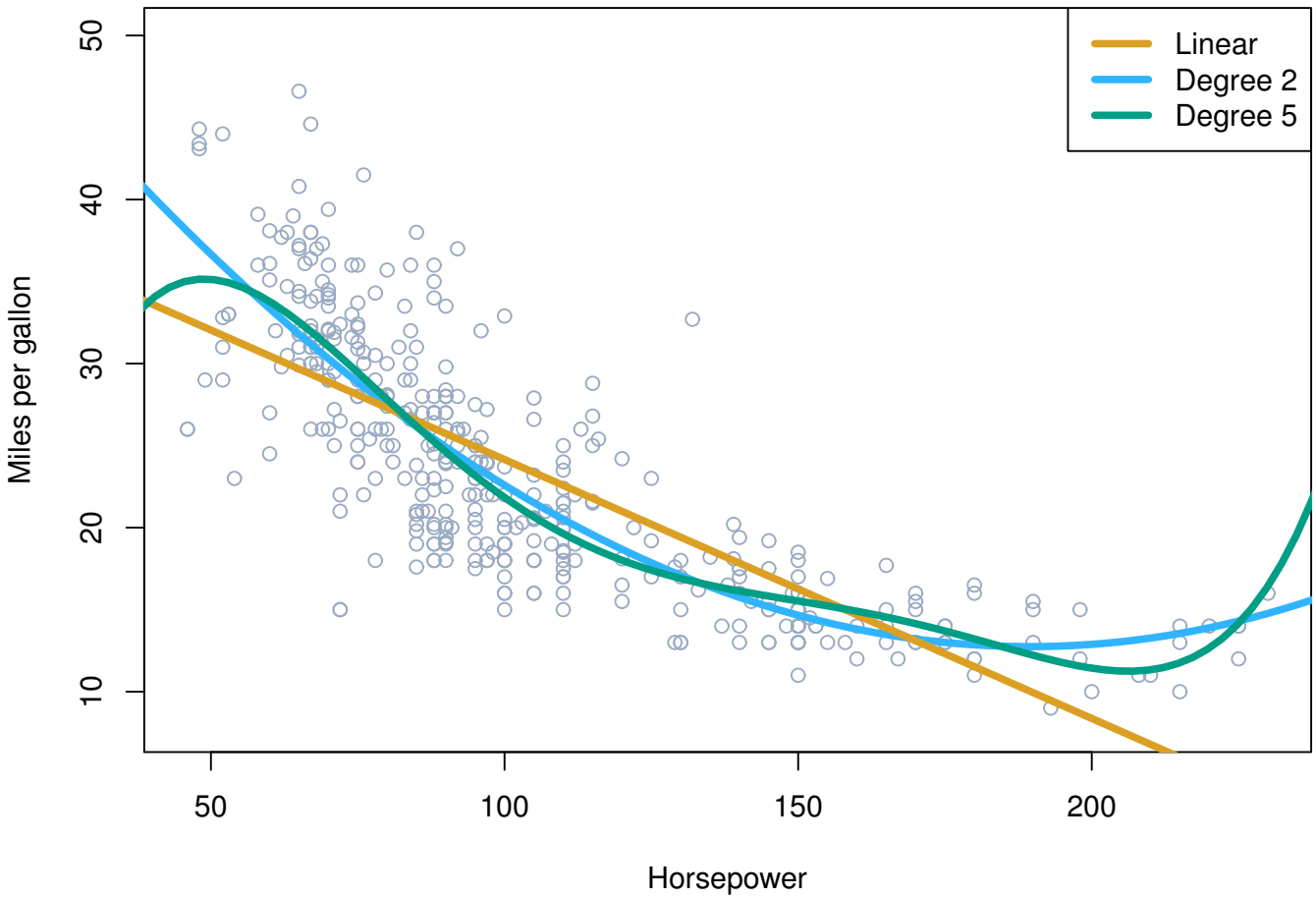
– **Example**: For a problem wtih only one feature, we have a set of point that look lying on a parabola, we use the regression $y = \beta_0 + \beta_1 x + \beta_2 x^2$



Can We Only Fit Flat Curves with Linear Regression?

– **Example**: Regressing Mile per Gallon in terms of the Horse Power

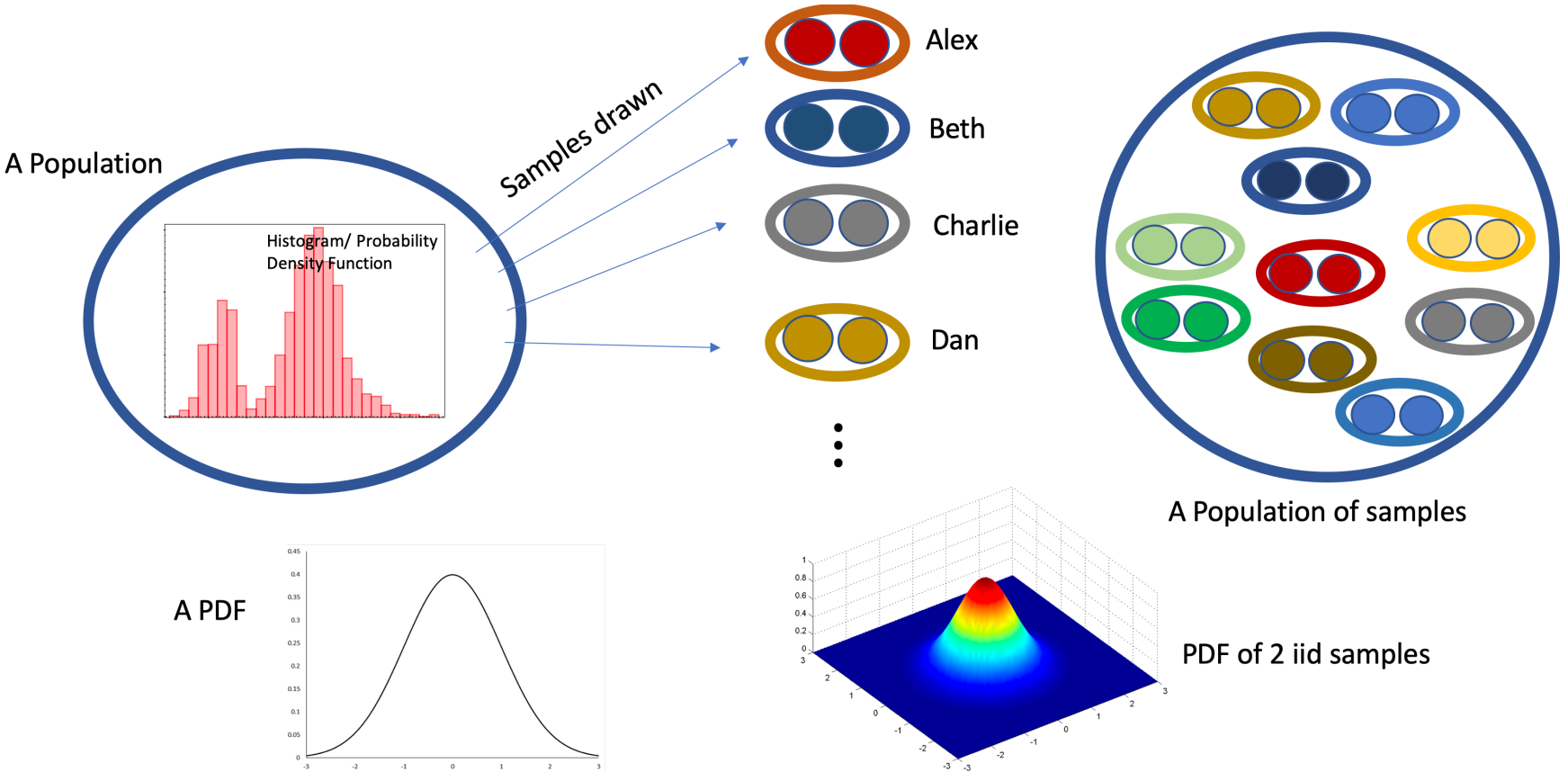
$$\texttt{mpg} = \beta_0 + \sum_{i=1}^{p} \beta_i (\texttt{horsepower})^i$$



Some Fundamental Notions and Tools in Machine Learning

Sampling Distribution

Sampling Distribution



- We often have a reference population (with some PDF or histogram) and would like to draw multiple samples to discover something about the population

Sampling Distribution

- A useful fact to derive the distribution for multiple independent samples is: *if* $x_1 \sim f_1(x)$, $x_2 \sim f_2(x)$, ..., $x_N \sim f_N(x)$ *are independent random variables, their joint PDF is*

$$f(x_1, x_2, \ldots, x_N) = f_1(x_1) f_2(x_2) \ldots f_N(x_N) = \prod_{i=1}^{N} f_i(x_i)$$

- **Example:** We take two independent samples x_1 and x_2 from a standard normal distribution. What is the joint PDF of x_1 and x_2?

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2}\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right) \end{aligned}$$

In-Class Exercise

- We take 10 samples from a normal distribution with mean μ and variance σ^2. What is the sampling distribution?

$$f(x_1, x_2, \ldots, x_N) = \ldots$$

Functions of Samples

- Not only we can talk about the joint distribution of the samples, we can also talk about the distribution of a function applied to the samples.
- Finding the distribution for functions of random variables is not a straightforward task.
- **Example**: If x_1 and x_2 have the joint PDF $f_{X,Y}(x, y)$, what is the pdf for $z = 3x + y$?

  Answer. $F_Z(z) = \mathbb{P}(3X + Y \leq z) = \int \int_{3x+y \leq z} f_{X,Y}(x, y) dx\ dy$ and then taking a derivative of the CDF $F_Z(z)$ to acquire the pdf $f_Z(z)$.
- But sometimes there are shortcuts. For example when we know what is the distribution for the sum of two random variables, and all we need to do is estimating the distribution parameters.

In-Class Exercise

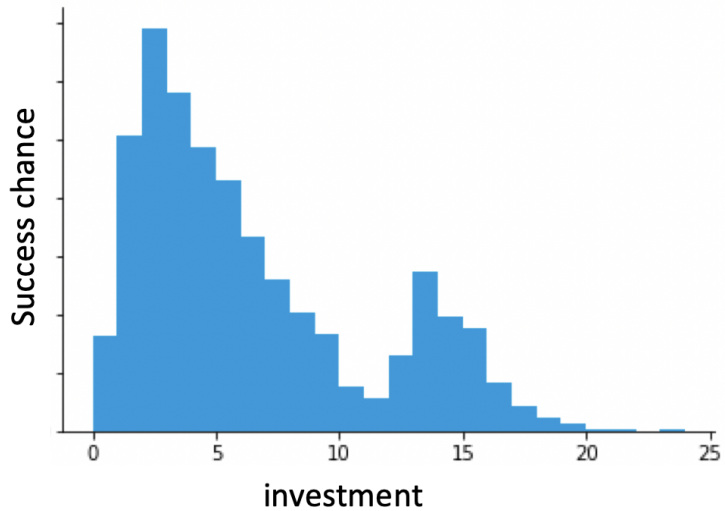
- **Example:** We have a normal distribution with mean 2 and variance 9. What is the distribution of of the sample mean acquired by averaging 4 independently drawn samples.
  Hard Way. We obtain the sampling distribution $f(x_1, x_2, x_3, x_4)$, which is the joint distribution, and then use the technique in the previous example to acquire the distribution of
  $z = (x_1 + x_2 + x_3 + x_4)/4$.
  Easy Way. Using the fact sheet from lecture 2, we know weighted sum of normal random variables is normally distributed, so all we need is to find the mean and the variance of
  $z = (x_1 + x_2 + x_3 + x_4)/4$.

$$\mathbb{E}(z) = 4 \times (2/4) = 2, \quad \text{var}(z) = 4 \times \frac{9}{16} = \frac{9}{4}.$$

  Therefore $z \sim \mathcal{N}(2, \frac{9}{4})$.
- See `https://en.wikipedia.org/wiki/Relationships_among_probability_distributions` for other shortcuts for the sum of independent random variables.

Brief Overview of Maximum Likelihood

– Say you are taking a risky investment and you know the following pdf corresponds to the succcess chance in terms of the investment. How much would you invest?



Brief Overview of Maximum Likelihood

- Maximum likelihood (ML) is a statistical estimation technique
- The main goal in ML is often estimating the parameters of the **reference** population from a set of samples
- Let $x_1, x_2, \cdots x_n$ be samples from a distribution with some unknown parameter θ and joint distribution

$$f(x_1, x_2, \cdots, x_n|\theta)$$

- The maximum likelihood estimate of θ based on the observations $\tilde{x}_1, \tilde{x}_2, \cdots \tilde{x}_n$ is

$$\theta_{ML} = \underset{\theta}{\text{argmax}} \; f(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n|\theta)$$

- When $x_1, x_2, \cdots x_n$ are i.i.d samples from a distribution $f(\cdot)$, then

$$f(x_1, x_2, \cdots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$$

Brief Overview of Maximum Likelihood

Example 1. We have a normal distribution $\mathcal{N}(\mu, 1)$ and we do not know μ. We take 5 independent samples from this distribution and the values turn out to be

$$\tilde{x}_1 = 2.5377,\ \tilde{x}_2 = 3.8339,\ \tilde{x}_3 = -0.2588,\ \tilde{x}_4 = 2.8622,\ \tilde{x}_5 = 2.3188,$$

what is the ML estimate of μ.

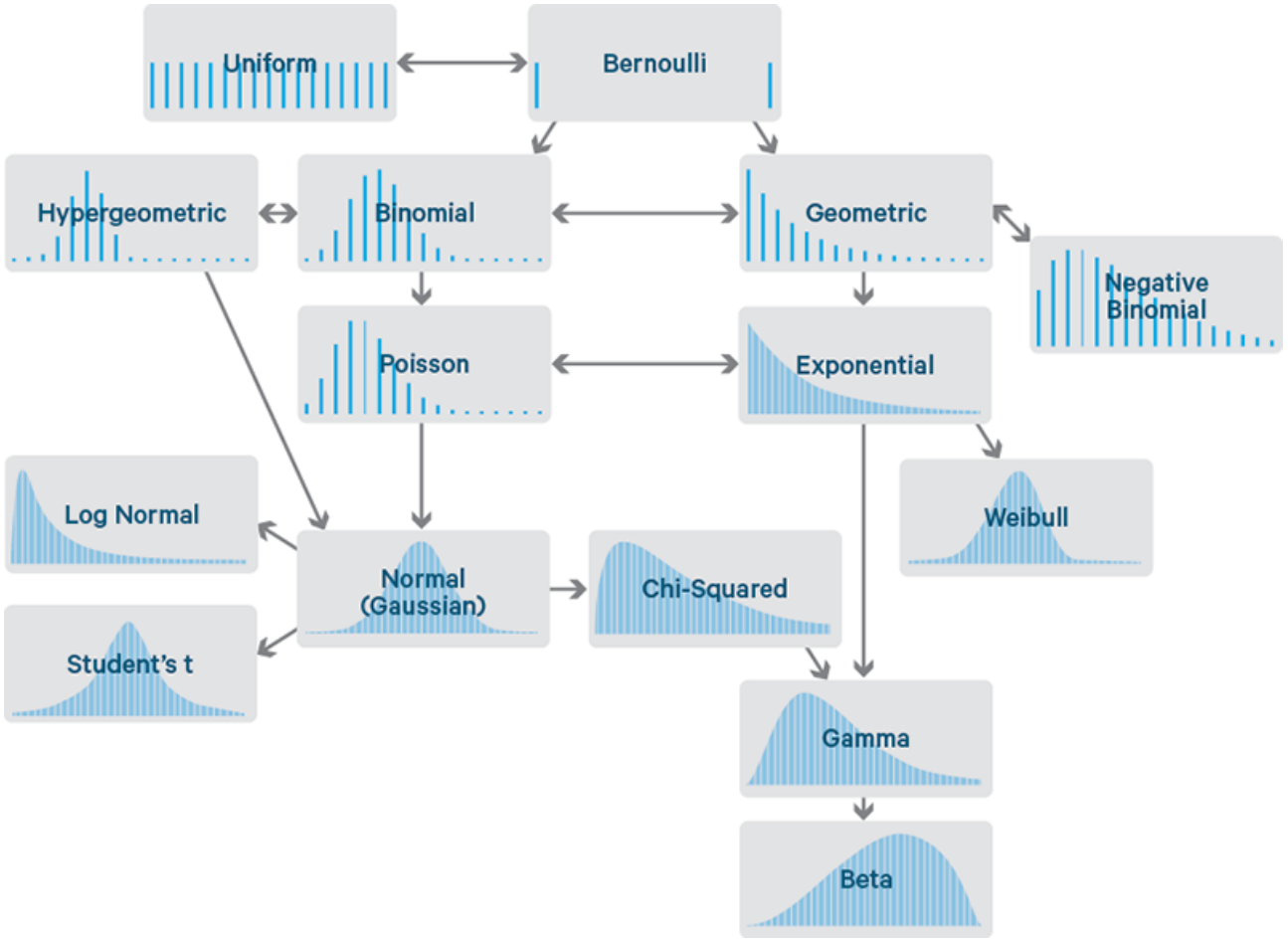
Solution. If we take 5 independent samples $x_1, x_2, \cdots x_5$ from a normal distribution $\mathcal{N}(\mu, 1)$, their joint distribution is

$$f(x_1, x_2, x_3, x_4, x_5|\mu) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_i - \mu)^2}{2}),$$

some basic calculus yields $\mu_{ML} = \frac{\tilde{x}_1 + \tilde{x}_2 + \cdots + \tilde{x}_5}{5} = 2.2587$ (why?)

Brief Overview of Maximum Likelihood

Example 2. We have a random generator which works as a black-box. Use the table below and a maximum likelihood approach to estimate the distribution and its parameters (see the MATLAB code).



[Figure Link]

Brief Overview of Maximum Likelihood

Example 3. We have a simple linear model in the form of $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. We pass the points $x_1, \ldots, x_n$ to the model and receive the independent random responses $y_1, \ldots, y_n$. Based on the observed samples, what is the ML estimate for β_0 and β_1?

Hint:

$$\begin{aligned} &\arg\max_{\beta_0, \beta_1} \quad f(y_1, \cdots y_n | X_1 = x_1, \cdots, X_n = x_n, \beta_0, \beta_1) \\ &= \arg\max_{\beta_0, \beta_1} \quad \log\left(f(y_1, \cdots y_n | X_1 = x_1, \cdots, X_n = x_n, \beta_0, \beta_1)\right) \end{aligned}$$

Important Note. Similar to the way we treated the RSS minimization, in ML we also need to set the derivative with respect to all the variables to zero. You will do this in the homework.

Brief Intro to Optimization

Setting the Derivative to Zero?

- We have been setting the derivative both for minimization and maximization. How do we know if the resulting point is a minimizer or a maximizer?
- What happens if we cannot solve the equation resulted by setting the derivative to zero?
- Normally distinguishing a minimizer from a maximizer requires the second derivative information
- In this lecture we will introduce methods that are for general minimization
- Also would cover a class of functions that a stationary point corresponds to a minimizer

Gradient Descent and I ts Variants

In-Class Exercise

- Find the minimizer of the function

$$\mathcal{C}(p_1, p_2) = (p_1 - p_2 - 3)^2 + p_2^2$$

Can We Always Find the Zero Easily?

- How about the minimizer of the function

$$\mathcal{C}(p_1, p_2) = (1 - p_1)^2 + (1 - p_2)^2 - 2\exp(-3p_1^2 - 3p_2^2)$$

Let's take a look at the function plot in Matlab.

Gradient Descent for Minimization

- We saw that our fitting and ML problems ultimately can be reduced to a minimization

$$\min_{\boldsymbol{p}} \ \mathcal{C}(\boldsymbol{p})$$

  where $\boldsymbol{p}$ includes all the unknown variables.
- Assuming $\boldsymbol{p} \in \mathbb{R}^L$, a numerical way of minimization is to start from a point $\boldsymbol{p}^{(0)}$ and iteratively perform the following steps

$$\boldsymbol{p}^{k+1} = \boldsymbol{p}^{(k)} - \eta \nabla \mathcal{C}\Big|_{\boldsymbol{p}=\boldsymbol{p}^{(k)}} \quad \text{where} \quad \nabla \mathcal{C} = \begin{pmatrix} \partial \mathcal{C}/\partial p_1 \\ \partial \mathcal{C}/\partial p_2 \\ \vdots \\ \partial \mathcal{C}/\partial p_L \end{pmatrix}$$

  parameter η is called the **learning rate**
- Larger learning rate does not necessarily mean faster solve
- Let's go through a simple example to see how gradient descent works (see the MATLAB code and the next slide)

Gradient Descent for Minimization

- Please refer to the MATLAB `gradientDescent.m` script
- Lets consider the very simple objective

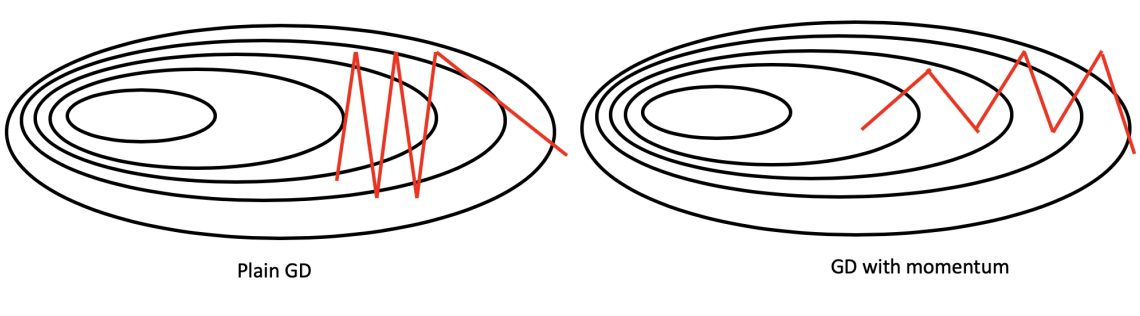
$$\mathcal{C}(p_1, p_2) = (1 - p_1)^2 + (1 - p_2)^2 - 2\exp(-3p_1^2 - 3p_2^2)$$

The gradient can be calculated as

$$\nabla\mathcal{C} = \begin{pmatrix} 2(p_1 - 1) + 12p_1\exp(-3p_1^2 - 3p_2^2) \\ 2(p_2 - 1) + 12p_2\exp(-3p_1^2 - 3p_2^2) \end{pmatrix}$$

- We can see that this objective has multiple local minimizers (two)
- Depending on where we start from we may land in either one
- A too small LR (learning rate) can make the minimization slow
- A too large LR can also make it slow or never converging!
- LR can affect which minimizer we converge to, but this is beyond our control

Gradient Descent with Momentum



- Momentum is a method that can dampen the gradient descent oscillations and accelerate it
- It can even help skipping shallow minima and land into deeper minima
- Gradient descent with **learning rate** η and **momentum** γ:

$$\boldsymbol{\theta}_{k+1} = \gamma \boldsymbol{\theta}_k + \eta \nabla \mathcal{C}(\boldsymbol{p}^k)$$
$$\boldsymbol{p}^{k+1} = \boldsymbol{p}^k - \boldsymbol{\theta}_{k+1}$$

- Again refer to the MATLAB code

Convex Functions

What Are Convex Functions?

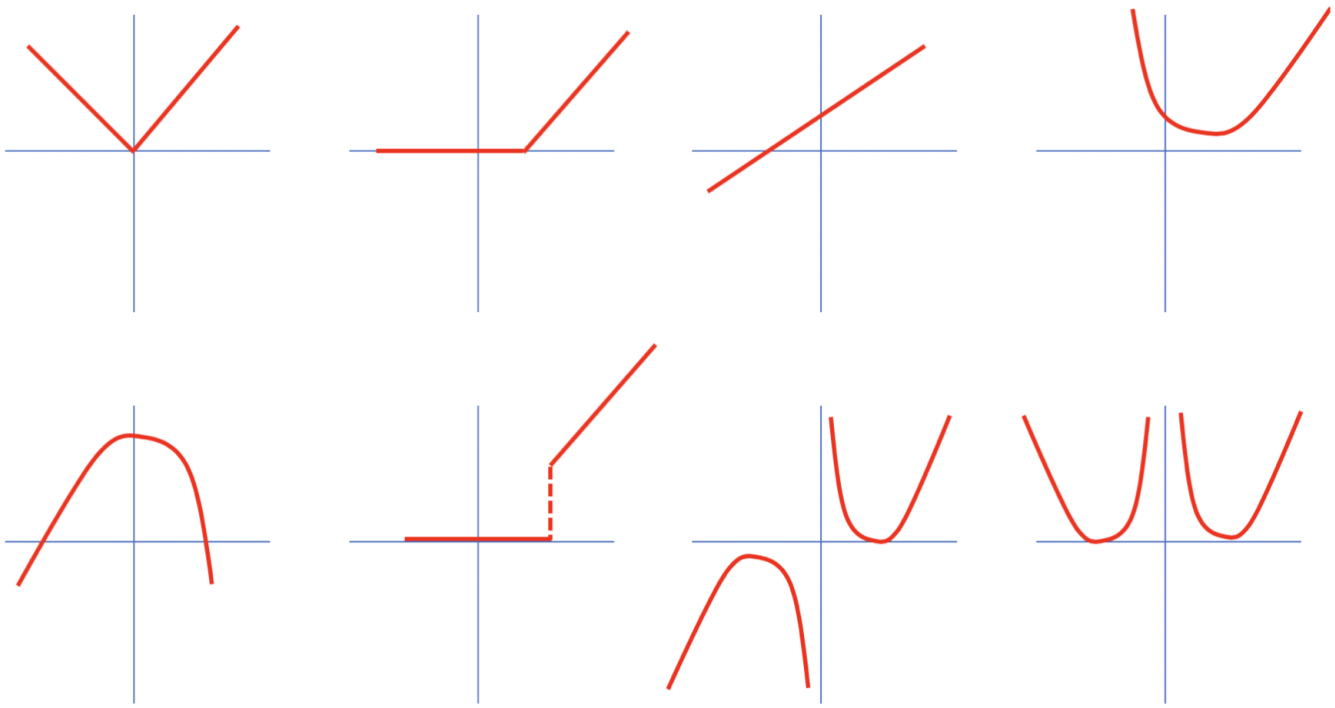
- Convex functions are a class of functions which have computationally attractive properties when it comes to minimization problems
- Suppose that $f(\boldsymbol{p}) : \mathbb{R}^n \to \mathbb{R}$, i.e., a function that operates on multiple variables (a vector) and produces a scalar as the output
- Function f is convex if for all $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in the domain, and $0 \leq \theta \leq 1$:

$$f(\theta \boldsymbol{p}_1 + (1-\theta)\boldsymbol{p}_2) \leq \theta f(\boldsymbol{p}_1) + (1-\theta) f(\boldsymbol{p}_2)$$

- Intuitively, a function is convex if the line segment between any two points on the graph of the function lies above (or just touching) the graph between the two points.

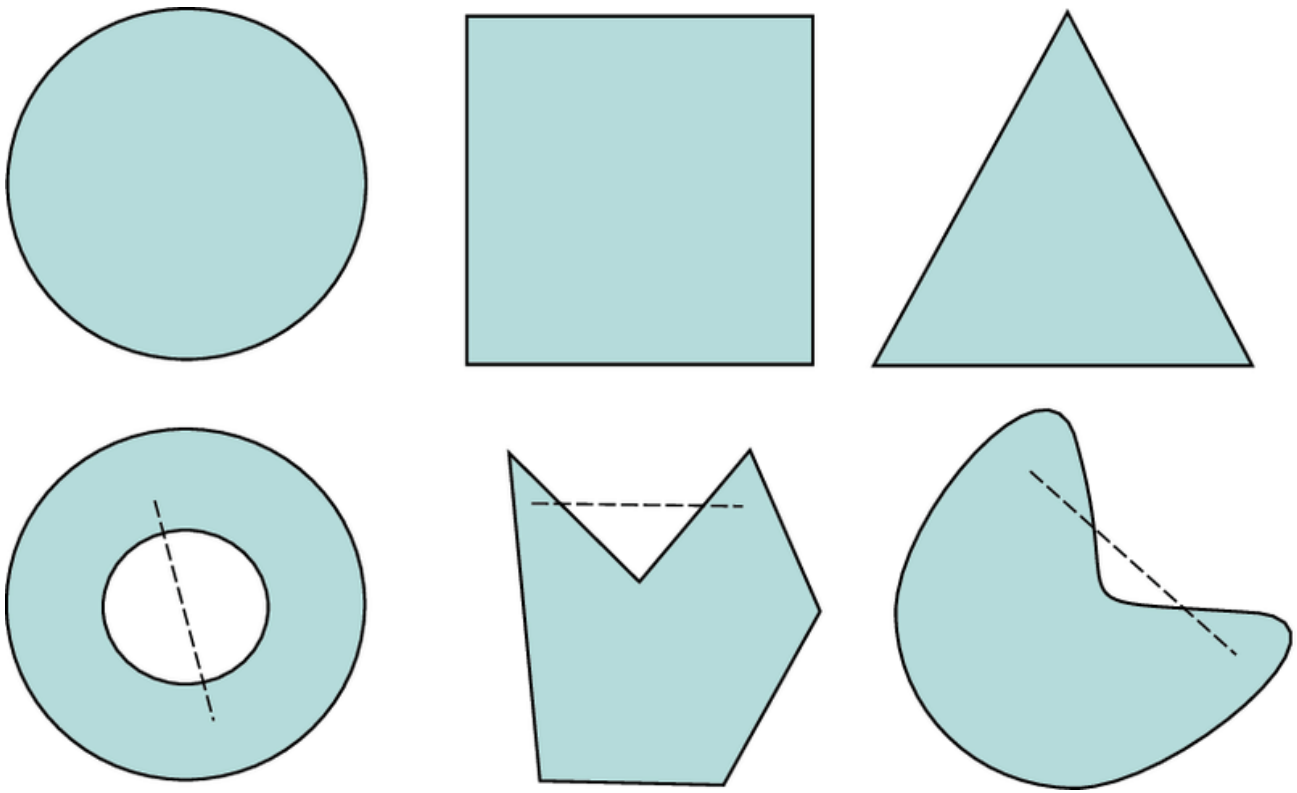
In-Class Exercise

- Identify convex functions:



Convex Sets

- A set is convex if the line connecting any two points of the set entirely stays within the set



More on Convexity

- In optimization theory, convex programs which are in the form

$$\underset{\boldsymbol{p}}{\text{minimize}} \quad f(\boldsymbol{p}) \qquad \text{subject to:} \qquad \boldsymbol{p} \in \text{a convex set}$$

  have very desirable computational properties
- For differentiable convex functions gradient descent always lands to a **global minimizer**
- A function that can be represented as the negative of a convex function is called **concave** function.
- Verifying the convexity in low dimensions, like 1 or 2, can be done visually. But for high dimensions we need to use the properties and definitions to show the convexity

More Properties and Examples

- One way to show the convexity is using the definition and showing that for all $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in the domain, and $0 \leq \theta \leq 1$:

$$f(\theta \boldsymbol{p}_1 + (1-\theta)\boldsymbol{p}_2) \leq \theta f(\boldsymbol{p}_1) + (1-\theta) f(\boldsymbol{p}_2).$$

More Properties and Examples

- Another way to show the convexity is using the following theorem: *A differentiable function is convex if for all* $\boldsymbol{p}_1$ *and* $\boldsymbol{p}_2$ *in the domain:*

$$f(\boldsymbol{p}_2) - f(\boldsymbol{p}_1) - \nabla f(\boldsymbol{p}_1)^\top (\boldsymbol{p}_2 - \boldsymbol{p}_1) \geq 0.$$

  (we got rid of θ)!

- **Example.** Show that the function $f(x, y) = (x + y)^2$ is convex.

- If $g(z)$ is convex, then g applied to a linear function is also convex, i.e., $g(\alpha_1 x_1 + \ldots + \alpha_n x_n)$ is convex.
- In other words, to show the convexity of $f(x, y) = (x + y)^2$ we only need to show that $f(z) = z^2$ is convex.

More Properties and Examples

- Yet, another way to show the convexity is using the following theorem: *A twice differentiable function is convex if at any point in the domain all the eigenvalues of the Hessian matrix are non-negative*. For $f(x_1, x_2, \ldots, x_n)$ the Hessian matrix is defined as

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \, \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \, \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \, \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \, \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \, \partial x_1} & \frac{\partial^2 f}{\partial x_n \, \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

- In the single variable case, a twice differentiable function $f(x)$ is convex if $f''(x)$ is non-negative for all the points in the domain.
- **Example.** Use this result to show that the function $f(x, y) = (x + y)^2$ is convex.

Questions?

References

https://www.alsharif.info/iom530, 2013.

J. Friedman, T. Hastie, and R. Tibshirani.
The elements of statistical learning.
Springer series in statistics, 2nd edition, 2009.

G. James, D. Witten, T. Hastie, and R. Tibshirani.
https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf, 2013.

G. James, D. Witten, T. Hastie, and R. Tibshirani.
https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/statistical_learning.pdf, 2013.

G. James, D. Witten, T. Hastie, and R. Tibshirani.
https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/linear_regression.pdf, 2013.

G. James, D. Witten, T. Hastie, and R. Tibshirani.
An introduction to statistical learning: with applications in R,