Deep Learning from Ground Up

More Linear Algebra

Akshay Badola
15MCPC15

School of Computer and Information Sciences
University of Hyderabad

December 2017

Overview

What is Statistics?

We've all heard the term

- So what is it?

What is Statistics?

We've all heard the term

- So what is it? Numbers based on observations?

What is Statistics?

We've all heard the term

- So what is it? Numbers based on observations? Summaries of numbers?

What is Statistics?

We've all heard the term

- So what is it? Numbers based on observations? Summaries of numbers?
- A statistic is any function of data.
- What is probability then, don't they both go together? And why do we need it for statistics?

What is Statistics?

We've all heard the term

- So what is it? Numbers based on observations? Summaries of numbers?
- A statistic is any function of data.
- What is probability then, don't they both go together? And why do we need it for statistics?
- Probability is the field of mathematics that deals with uncertain events. We need it for statistics because data is **inherently** uncertain.
- I hope we all know the basics of probability because we won't repeat it here.

What is Statistics?

We've all heard the term

- So what is it? Numbers based on observations? Summaries of numbers?
- A statistic is any function of data.
- What is probability then, don't they both go together? And why do we need it for statistics?
- Probability is the field of mathematics that deals with uncertain events. We need it for statistics because data is **inherently** uncertain.
- I hope we all know the basics of probability because we won't repeat it here.
- We will enumerate some identities for multivariate probability distributions.

Distribution

What's a distribution now?

- A probability distribution is some function which assigns a probability measure to a set Ω.
- For every member in the set Ω, it gives us a value (between 0 and 1).
- For example, the function $f(x) = e^{a+bx}$, for $-1 \leq x \leq 1$ assigns a number to each value in the set {-1, 1}, but the numbers do not sum to one.
- However $P(x) = \frac{e^{a+bx}}{\int_{-1}^{1} e^{a+by}}$ does sum (integrate) to one for $-1 \leq x \leq 1$.

Distribution

What's a distribution now?

- A probability distribution is some function which assigns a probability measure to a set Ω.
- For every member in the set Ω, it gives us a value (between 0 and 1).
- For example, the function $f(x) = e^{a+bx}$, for $-1 \leq x \leq 1$ assigns a number to each value in the set {-1, 1}, but the numbers do not sum to one.
- However $P(x) = \frac{e^{a+bx}}{\int_{-1}^{1} e^{a+by}}$ does sum (integrate) to one for $-1 \leq x \leq 1$.
- What's the value of $\int_y e^{a+by}$, for $-1 \leq y \leq 1$.
- The denominator is the called the *Normalizing Function* or sometimes the *Partition Function* (the term comes from statistical physics, google it!).

Probability and Uncertainty

Dealing with uncertainty

- "Probability does not exist!", wrote De Finetti in the preface of his classic[1] *"Theory of Probability"*[1]

[1] "De Finetti predicts in these volumes that we shall all be Bayesians by 2020." - Adrian Smith

Probability and Uncertainty

Dealing with uncertainty

- "Probability does not exist!", wrote De Finetti in the preface of his classic[1] *"Theory of Probability"*[1]
- He meant it in the sense that there's no such thing as *objective probability*. And that it is simply a tool we use to deal with uncertainty.
- Regardless of what the actual process is, be it deterministic or non-deterministic, we deal with it by ascribing that uncertainty to our knowledge and actions.

[1] "De Finetti predicts in these volumes that we shall all be Bayesians by 2020." - Adrian Smith

Probability and Uncertainty

Dealing with uncertainty

- "Probability does not exist!", wrote De Finetti in the preface of his classic[1] *"Theory of Probability"*[1]
- He meant it in the sense that there's no such thing as *objective probability*. And that it is simply a tool we use to deal with uncertainty.
- Regardless of what the actual process is, be it deterministic or non-deterministic, we deal with it by ascribing that uncertainty to our knowledge and actions.
- We will talk of probability while dealing exclusively with data and there can always be some error in gathering data, or data itself may not represent the truth (that which we're seeking).
- We say in such cases that there is **inherent noise** in the observations and we refer to it as *observational noise*. It simply exists because the world can't be measured accurately enough.

[1] "De Finetti predicts in these volumes that we shall all be Bayesians by 2020." - Adrian Smith

Probability and Models

Modelling the World

- We'll talk about what it means to be bayesian in a while, but first we'll mention what a model is.
- A model can be thought of as a mathematical formulation of an event or set of events, which aims to provide information on it or them.
- A model therefore aims to tell us *something* about those events. What that *something* is, depends on the events and the model.
- For instance there can be a model of heat distribution across a room and it may not necessarily be a probability distribution.

Probability and Models

Modelling the World

- We'll talk about what it means to be bayesian in a while, but first we'll mention what a model is.
- A model can be thought of as a mathematical formulation of an event or set of events, which aims to provide information on it or them.
- A model therefore aims to tell us *something* about those events. What that *something* is, depends on the events and the model.
- For instance there can be a model of heat distribution across a room and it may not necessarily be a probability distribution.
- There are two questions we'll always want to know:
  1. How accurate is our model?
  2. How can we make a better model?
- Accuracy can be estimated perhaps by measuring some aspects of the world (recall assignments and Linear Algebra)
- But how do we make one? Well it depends on what we're trying to model.

Statistical Models

Modelling Data

- If we're trying to model the trajectory of an object, we'd be better off using the laws of physics.
- When we'll talk of models we'll mean Statistical Models, that is, models which try to find some functions of data. What does it mean to find a function of data?

Statistical Models

Modelling Data

- If we're trying to model the trajectory of an object, we'd be better off using the laws of physics.
- When we'll talk of models we'll mean Statistical Models, that is, models which try to find some functions of data. What does it mean to find a function of data?
- In the most general sense, we want some function $f : \text{Some Data} \rightarrow \text{Rules}$, such that $f(\text{Rules}) = \text{All Similar Data}$
- In short we want to find the laws governing the data. Since we can't be absolutely certain of either the nature of the data (observational uncertainty) or the nature of the model (systematic error), there's always some uncertainty both in the function we model and data provided by that model.
- This is different than the *bias/variance dilemma* we hear of. (Which we'll get to in time.)

Statistical Models

Choices (Assumptions) we make

- Getting back to finding the function, the simplest way to find it is to assume that the data was generated by some function with certain characteristics and find those characteristics.

Statistical Models

Choices (Assumptions) we make

- Getting back to finding the function, the simplest way to find it is to assume that the data was generated by some function with certain characteristics and find those characteristics.
- We say that the *data was generated by the distribution* and we try to find the *parameters* of the distribution.
- The simplest assumption we can make in case of discrete events is that all events are equally likely. This is known as the *Uniform* distribution.
- What is the distribution for a sequence of events, with no turn of a sequence affecting the other and both events being equally likely?

Statistical Models

Choices (Assumptions) we make

- Getting back to finding the function, the simplest way to find it is to assume that the data was generated by some function with certain characteristics and find those characteristics.
- We say that the *data was generated by the distribution* and we try to find the *parameters* of the distribution.
- The simplest assumption we can make in case of discrete events is that all events are equally likely. This is known as the *Uniform* distribution.
- What is the distribution for a sequence of events, with no turn of a sequence affecting the other and both events being equally likely? The *Binomial* of course!
- What would be a distribution for the assumption (for given data of real numbers) there's only one number and it is surely within ± 0.5 of it? It does feel very symmetric about the uncertainty though, right?

Statistical Models

Choices (Assumptions) we make

- Getting back to finding the function, the simplest way to find it is to assume that the data was generated by some function with certain characteristics and find those characteristics.
- We say that the *data was generated by the distribution* and we try to find the *parameters* of the distribution.
- The simplest assumption we can make in case of discrete events is that all events are equally likely. This is known as the *Uniform* distribution.
- What is the distribution for a sequence of events, with no turn of a sequence affecting the other and both events being equally likely? The *Binomial* of course!
- What would be a distribution for the assumption (for given data of real numbers) there's only one number and it is surely within ± 0.5 of it? It does feel very symmetric about the uncertainty though, right?
- It's the normal distribution or the *Gaussian*.

Statistical Models

Odds and Ends

We'll define some terms and notations:

- We'll denote by $\mathbb{E}(x)$ (or sometimes $\mu(x)$ and call it mean) the expectation of a random variable x, and it's value will be calculated as $\mathbb{E}(x) = \sum xf(x)$ or $\int_\Omega xf(x)dx$.

Statistical Models

Odds and Ends

We'll define some terms and notations:

- We'll denote by $\mathbb{E}(x)$ (or sometimes $\mu(x)$ and call it mean) the expectation of a random variable x, and it's value will be calculated as $\mathbb{E}(x) = \sum xf(x)$ or $\int_{\Omega} xf(x)dx$.
- $\mathbb{E}(x)$ is linear! $\mathbb{E}(x+y) = \mathbb{E}(x) + \mathbb{E}(y)$ for two rv's x and y.
- $\mathbb{E}(x)$ is called the first moment of x and the second moment (around $\mathbb{E}(x)$ or μ. We can leave x out if its clear from context) is, $\mathbb{E}(x-\mu)^2 = \sum \mathbb{E}(x^2) - \mathbb{E}^2(x)$ (prove. Homework) or, $\int_{\Omega} (x-\mu)^2 f(x)dx$, also called variance (around the mean) σ^2

Statistical Models

Odds and Ends

We'll define some terms and notations:

- We'll denote by $\mathbb{E}(x)$ (or sometimes $\mu(x)$ and call it mean) the expectation of a random variable x, and it's value will be calculated as $\mathbb{E}(x) = \sum x f(x)$ or $\int_{\Omega} x f(x) dx$.
- $\mathbb{E}(x)$ is linear! $\mathbb{E}(x + y) = \mathbb{E}(x) + \mathbb{E}(y)$ for two rv's x and y.
- $\mathbb{E}(x)$ is called the first moment of x and the second moment (around $\mathbb{E}(x)$ or μ. We can leave x out if its clear from context) is, $\mathbb{E}(x - \mu)^2 = \sum \mathbb{E}(x^2) - \mathbb{E}^2(x)$ (prove. Homework) or, $\int_{\Omega} (x - \mu)^2 f(x) dx$, also called variance (around the mean) σ^2
- In general for an arbitrary function $g, \mathbb{E}(g)$ can be defined by simply substituting g for x. Note that $\mathbb{E}$ depends on the variable and if g has as its domain variables other than x then $\mathbb{E}$ has to be specified, e.g., $\mathbb{E}_x(g) = \int (x^2 + y^2) f(x) dx$, for $g(x, y) = x^2 + y^2 =?$ (Classwork! hint: use linearity of Expectation)

Parametric Models

Back to Models!

- But we were trying to find some *parameters* of some models.

Parametric Models

Back to Models!

- But we were trying to find some *parameters* of some models. If we assume that the data is generated by some distribution which is characterized by some function $f(x_1, x_2, x_3...)$ then $\{x_1, x_2, x_3...\}$ are the parameters of the distribution.
- In our case, we'll have to find them from data, hence they are a function of data (mean, variance both are functions of data)

Parametric Models

Back to Models!

- But we were trying to find some *parameters* of some models. If we assume that the data is generated by some distribution which is characterized by some function $f(x_1, x_2, x_3 ...)$ then $\{x_1, x_2, x_3 ...\}$ are the parameters of the distribution.
- In our case, we'll have to find them from data, hence they are a function of data (mean, variance both are functions of data)
- We say we perform *inference* for the parameters. We won't go into the details of inference here, however we'll talk about some concepts as required.
- Revisiting the Gaussian, from the formula $\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{2} \frac{(x-\mu)}{\sigma^2}}$ we see that the distribution is completely specified by μ and σ^2, hence, we say that the Gaussian is parametrized by its mean and variance (also called location and scale in some contexts).
- In short we make an assumption that the data was generated by some distribution (say the Gaussian) and we try to estimate its parameters (μ and σ^2) from the data we have.

Likelihood

How do we find the parameters given the data?

- There can be a few ways, we'll discuss only one. Likelihood.
- A History Lesson! Fisher, Likelihood and Bayes (or Laplace).

Likelihood

How do we find the parameters given the data?

- There can be a few ways, we'll discuss only one. Likelihood.
- A History Lesson! Fisher, Likelihood and Bayes (or Laplace).
- The Likelihood is some function which measures the plausibility of the parameters[2] of the distribution we're interested in. It is defined to be some function of θ, proportional to the model density, $\mathcal{L}(\theta) = \mathcal{L}(\boldsymbol{\theta}|x) = cf(x, \boldsymbol{\theta})$.

Likelihood

How do we find the parameters given the data?

- There can be a few ways, we'll discuss only one. Likelihood.
- A History Lesson! Fisher, Likelihood and Bayes (or Laplace).
- The Likelihood is some function which measures the plausibility of the parameters[2] of the distribution we're interested in. It is defined to be some function of θ, proportional to the model density, $\mathcal{L}(\theta) = \mathcal{L}(\boldsymbol{\theta}|x) = cf(x, \boldsymbol{\theta})$.
- For instance, for the one dimensional Gaussian, suppose we want to find out its parameters, μ and σ^2, how do we start?
- Well first we form the likelihood function

Likelihood

How do we find the parameters given the data?

- There can be a few ways, we'll discuss only one. Likelihood.
- A History Lesson! Fisher, Likelihood and Bayes (or Laplace).
- The Likelihood is some function which measures the plausibility of the parameters[2] of the distribution we're interested in. It is defined to be some function of θ, proportional to the model density, $\mathcal{L}(\theta) = \mathcal{L}(\boldsymbol{\theta}|x) = cf(x, \boldsymbol{\theta})$.
- For instance, for the one dimensional Gaussian, suppose we want to find out its parameters, μ and σ^2, how do we start?
- Well first we form the likelihood function $\mathcal{L}(\theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2}(\frac{x-\theta_1}{\theta_2})^2}$ where $\boldsymbol{\theta} = (\theta_1, \theta_2)$

Likelihood

How do we find the parameters given the data?

- There can be a few ways, we'll discuss only one. Likelihood.
- A History Lesson! Fisher, Likelihood and Bayes (or Laplace).
- The Likelihood is some function which measures the plausibility of the parameters[2] of the distribution we're interested in. It is defined to be some function of θ, proportional to the model density, $\mathcal{L}(\theta) = \mathcal{L}(\boldsymbol{\theta}|x) = cf(x, \boldsymbol{\theta})$.
- For instance, for the one dimensional Gaussian, suppose we want to find out its parameters, μ and σ^2, how do we start?
- Well first we form the likelihood function $\mathcal{L}(\theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2}(\frac{x-\theta_1}{\theta_2})^2}$ where $\boldsymbol{\theta} = (\theta_1, \theta_2)$
- So, if we have some data $x_i = X$ and we assume that each data point is drawn uniformly from the population and is independent of the other. We say $\mathcal{L}(\theta_1, \theta_2|X) = ?$ (What would independence imply?)

Likelihood

How do we find the parameters given the data?

- There can be a few ways, we'll discuss only one. Likelihood.
- A History Lesson! Fisher, Likelihood and Bayes (or Laplace).
- The Likelihood is some function which measures the plausibility of the parameters[2] of the distribution we're interested in. It is defined to be some function of θ, proportional to the model density, $\mathcal{L}(\theta) = \mathcal{L}(\boldsymbol{\theta}|x) = cf(x, \boldsymbol{\theta})$.
- For instance, for the one dimensional Gaussian, suppose we want to find out its parameters, μ and σ^2, how do we start?
- Well first we form the likelihood function $\mathcal{L}(\theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2}(\frac{x-\theta_1}{\theta_2})^2}$ where $\boldsymbol{\theta} = (\theta_1, \theta_2)$
- So, if we have some data $x_i = X$ and we assume that each data point is drawn uniformly from the population and is independent of the other. We say $\mathcal{L}(\theta_1, \theta_2|X) = \prod \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2}(\frac{x_i-\theta_1}{\theta_2})^2}$

Likelihood

Maximum Likelihood

- It's convenient to use the logarithms for easier calculations, so this becomes *Log Likelihood*
  $\log \mathcal{L}(\theta_1, \theta_2 | X) = \sum(-\frac{1}{2} \log 2\pi\theta_2 - \frac{1}{2}(\frac{x_i - \theta_1}{\theta_2})^2)$

Likelihood

Maximum Likelihood

- It's convenient to use the logarithms for easier calculations, so this becomes *Log Likelihood*

$$\begin{aligned}\log\mathcal{L}(\theta_1,\theta_2|X) &= \sum(-\tfrac{1}{2}\log 2\pi\theta_2 - \tfrac{1}{2}(\tfrac{x_i-\theta_1}{\theta_2})^2) \\ &= -\tfrac{1}{2}\sum(\log 2\pi\theta_2 - (\tfrac{x_i-\theta_1}{\theta_2})^2)\end{aligned}$$

Likelihood

Maximum Likelihood

- It's convenient to use the logarithms for easier calculations, so this becomes *Log Likelihood*
$$\begin{aligned}\log \mathcal{L}(\theta_1, \theta_2|X) &= \sum(-\tfrac{1}{2}\log 2\pi\theta_2 - \tfrac{1}{2}(\tfrac{x_i-\theta_1}{\theta_2})^2) \\ &= -\tfrac{1}{2}\sum(\log 2\pi\theta_2 - (\tfrac{x_i-\theta_1}{\theta_2})^2)\end{aligned}$$
- How do we maximize?

Likelihood

Maximum Likelihood

- It's convenient to use the logarithms for easier calculations, so this becomes *Log Likelihood*
  $$\begin{aligned}\log \mathcal{L}(\theta_1, \theta_2|X) &= \sum(-\tfrac{1}{2}\log 2\pi\theta_2 - \tfrac{1}{2}(\tfrac{x_i-\theta_1}{\theta_2})^2) \\ &= -\tfrac{1}{2}\sum(\log 2\pi\theta_2 - (\tfrac{x_i-\theta_1}{\theta_2})^2)\end{aligned}$$
- How do we maximize?

  We just equate the gradient to zero.
  $$\hat{\theta_1} = \frac{\partial \log \mathcal{L}(\theta_1,\theta_2|X)}{\partial\theta_1} = -\sum(\frac{x_i-\theta_1}{\theta_2}) = 0 \Rightarrow \hat{\theta_1} = \sum\frac{x_i}{n} = \bar{x}$$
- So $\hat{\theta_1} = \hat{\mu}$. Similarly we can show that $\hat{\theta_2} = \hat{\sigma^2} = \sum\frac{(x_i-\bar{x})^2}{n}$

Likelihood

Maximum Likelihood

- It's convenient to use the logarithms for easier calculations, so this becomes *Log Likelihood*
  $\log \mathcal{L}(\theta_1, \theta_2|X) = \sum(-\frac{1}{2}\log 2\pi\theta_2 - \frac{1}{2}(\frac{x_i - \theta_1}{\theta_2})^2)$
  $= -\frac{1}{2}\sum(\log 2\pi\theta_2 - (\frac{x_i - \theta_1}{\theta_2})^2)$
- How do we maximize?
  We just equate the gradient to zero.
  $\hat{\theta_1} = \frac{\partial \log \mathcal{L}(\theta_1, \theta_2|X)}{\partial \theta_1} = -\sum(\frac{x_i - \theta_1}{\theta_2}) = 0 \Rightarrow \hat{\theta_1} = \sum \frac{x_i}{n} = \bar{x}$
- So $\hat{\theta_1} = \hat{\mu}$. Similarly we can show that $\hat{\theta_2} = \hat{\sigma^2} = \sum \frac{(x_i - \bar{x})^2}{n}$
- So, say the data is $\{f(2,8), f(3,5), f(6,4), f(10,2)\}$, where $f(a,b) = \underbrace{a, a, ..., a}_{\text{b times}}$, if we assume it to be drawn from a Gaussian distribution, we can immediately arrive at the maximum likelihood estimate $\hat{\mu} = 3.8$ and $\hat{\sigma^2} = ?$ Homework!

Likelihood

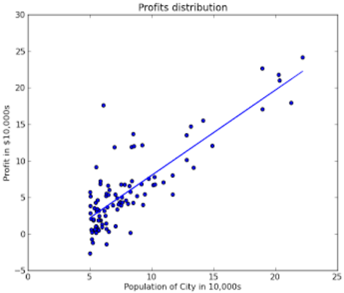
A bit more about the Likelihood

-

Linear Models

The Simplest Case

- A Linear Model makes the assumption that the data points are linearly related.



Linear Models

The Simplest Case contd.

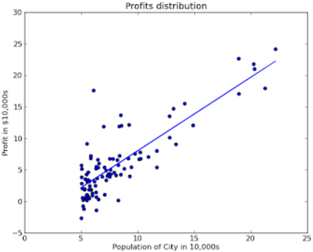
- Here we're talking about *Regression*. Regression means that we want to predict a numerical value for a *response* or *dependent* or *target* variable y_i corresponding to a set of *independent* or *input* or *explanatory* variables x_i.
- Although we'll mostly talk about one dimensional x, y, or perhaps two dimensional $\mathbf{x}$ and one dimensional y, both the input and the target variables can exist in multiple dimensions.

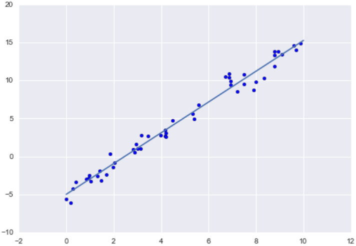
Linear Models

The Simplest Case contd.

- Here we're talking about *Regression*. Regression means that we want to predict a numerical value for a *response* or *dependent* or *target* variable y_i corresponding to a set of *independent* or *input* or *explanatory* variables x_i.
- Although we'll mostly talk about one dimensional x, y, or perhaps two dimensional **x** and one dimensional y, both the input and the target variables can exist in multiple dimensions.

Which is a better fit?





Formulation

What is Linear anyway?

- We've already seen the concept of Linearity and its importance. Linear Models are the simplest models and make quite strong assumptions regarding the data.
- We all know the equation of the straight line $y = mx + c$

Formulation

What is Linear anyway?

- We've already seen the concept of Linearity and its importance. Linear Models are the simplest models and make quite strong assumptions regarding the data.
- We all know the equation of the straight line or equivalently $y = wx + b$

Formulation

What is Linear anyway?

- We've already seen the concept of Linearity and its importance. Linear Models are the simplest models and make quite strong assumptions regarding the data.
- We all know the equation of the straight line or equivalently $y = wx + b$
- In some cases you'll also see $y_i = \beta x_i + \epsilon_i$, where $\beta = \beta_0, \beta_1$ and $x = x_0, x_1$
- It's the same formulation as previous for a straight line except X is now what's referred to as *augmented* vector and ϵ is the observational noise which can never (realistically) be zero.

Formulation

What is Linear anyway?

- We've already seen the concept of Linearity and its importance. Linear Models are the simplest models and make quite strong assumptions regarding the data.
- We all know the equation of the straight line or equivalently $y = wx + b$
- In some cases you'll also see $y_i = \beta x_i + \epsilon_i$, where $\beta = \beta_0, \beta_1$ and $x = x_0, x_1$
- It's the same formulation as previous for a straight line except X is now what's referred to as *augmented* vector and ϵ is the observational noise which can never (realistically) be zero.
- So for a set of data points (y_i, x_i):

$$Y \equiv \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ or, } Y = X\beta + \epsilon$$

- Now all we have to do is to find β, right?

Formulation

What is Linear anyway?

- We've already seen the concept of Linearity and its importance. Linear Models are the simplest models and make quite strong assumptions regarding the data.
- We all know the equation of the straight line or equivalently $y = wx + b$
- In some cases you'll also see $y_i = \beta x_i + \epsilon_i$, where $\beta = \beta_0, \beta_1$ and $x = x_0, x_1$
- It's the same formulation as previous for a straight line except X is now what's referred to as *augmented* vector and ϵ is the observational noise which can never (realistically) be zero.
- So for a set of data points (y_i, x_i):

$$Y \equiv \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ or, } Y = X\beta + \epsilon$$

- Now all we have to do is to find β, right? That depends.

Formulation

It's again in the assumptions we make

- In most cases, you'll see a simple Ordinary Least Squares (OLS) Estimate.
- The simples perspective is to go with our assumptions and try to find values of parameters which give the best results.
- For a straight line assumption, we can try to find a line which is the least average distance away from the data points. This is the geometrical perspective.

Formulation

It's again in the assumptions we make

- In most cases, you'll see a simple Ordinary Least Squares (OLS) Estimate.
- The simples perspective is to go with our assumptions and try to find values of parameters which give the best results.
- For a straight line assumption, we can try to find a line which is the least average distance away from the data points. This is the geometrical perspective.
- $\sum d_i^2 = \sum (y_i - \beta x_i)^2$, where x_k, β are as defined above.
- For the minimal distance, we differentiate w.r.t. β, so, $\frac{\partial d_i^2}{\partial \beta} = 2 \sum x_i (y_i - \beta x_i)$,

Formulation

It's again in the assumptions we make

- In most cases, you'll see a simple Ordinary Least Squares (OLS) Estimate.
- The simples perspective is to go with our assumptions and try to find values of parameters which give the best results.
- For a straight line assumption, we can try to find a line which is the least average distance away from the data points. This is the geometrical perspective.
- $\sum d_i^2 = \sum (y_i - \beta x_i)^2$, where x_k, β are as defined above.
- For the minimal distance, we differentiate w.r.t. β, so, $\frac{\partial d_i^2}{\partial \beta} = 2 \sum x_i(y_i - \beta x_i)$, or, $\frac{1}{n} \sum^n \beta = \frac{1}{n} \sum^n y_i / x_i$

Formulation

It's again in the assumptions we make

- In most cases, you'll see a simple Ordinary Least Squares (OLS) Estimate.
- The simples perspective is to go with our assumptions and try to find values of parameters which give the best results.
- For a straight line assumption, we can try to find a line which is the least average distance away from the data points. This is the geometrical perspective.
- $\sum d_i^2 = \sum (y_i - \beta x_i)^2$, where x_k, β are as defined above.
- For the minimal distance, we differentiate w.r.t.

  β, so, $\frac{\partial d_i^2}{\partial \beta} = 2 \sum x_i(y_i - \beta x_i)$, or, $\frac{1}{n} \sum^n \beta = \frac{1}{n} \sum^n y_i / x_i$

  Or, $\beta = \frac{1}{n} \sum^n y_i / x_i$, So simple!

Formulation

It's again in the assumptions we make

- In most cases, you'll see a simple Ordinary Least Squares (OLS) Estimate.
- The simples perspective is to go with our assumptions and try to find values of parameters which give the best results.
- For a straight line assumption, we can try to find a line which is the least average distance away from the data points. This is the geometrical perspective.
- $\sum d_i^2 = \sum (y_i - \beta x_i)^2$, where x_k, β are as defined above.
- For the minimal distance, we differentiate w.r.t. β, so, $\frac{\partial d_i^2}{\partial \beta} = 2 \sum x_i(y_i - \beta x_i)$, or, $\frac{1}{n} \sum^n \beta = \frac{1}{n} \sum^n y_i/x_i$
  Or, $\beta = \frac{1}{n} \sum^n y_i/x_i$, So simple!
- The response variable can now be written as $y_{new} = x_{new}\beta$
- However if we were to approach it from a statistical perspective, we would like to quantify the uncertainty in our estimated parameter.

Formulation

Slightly Deeper

- We make a slight augmentation and an assumption.
- First notice that:

$$X\beta = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = X\beta \text{ for } X \text{ of degree } 2$$

- The same formulation suffices for all polynomial models as the model is still *linear in the parameters*.

Formulation

Slightly Deeper

- We make a slight augmentation and an assumption.
- First notice that:

$$X\beta = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = X\beta \text{ for } X \text{ of degree } 2$$

- The same formulation suffices for all polynomial models as the model is still *linear in the parameters*.
- So we make an assumption on β: $\beta \sim \mathcal{N}(\beta, \mathbf{V}_\beta)$, with this we've gone Bayesian! (More on that later)
- What it means is that the parameter may have some value but we cannot really know it. The variance quantifies the uncertainty.

Formulation

Slightly Deeper

- We make a slight augmentation and an assumption.
- First notice that:

$$X\beta = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = X\beta \text{ for } X \text{ of degree } 2$$

- The same formulation suffices for all polynomial models as the model is still *linear in the parameters*.
- So we make an assumption on β: $\beta \sim \mathcal{N}(\beta, \mathbf{V}_\beta)$, with this we've gone Bayesian! (More on that later)
- What it means is that the parameter may have some value but we cannot really know it. The variance quantifies the uncertainty.

- So, again our model is $Y = X\beta + \epsilon,\ Y \sim \mathcal{N}(\mu, \mathbf{I}\sigma^2), \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$
  For OLS estimate we have to minimize $(Y - X\beta)^2$, again differentiating w.r.t. β we get:
  $-X(Y - X\beta) = 0$

- So, again our model is $Y = X\beta + \epsilon,\ Y \sim \mathcal{N}(\mu, \mathbf{I}\sigma^2), \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$
  For OLS estimate we have to minimize $(Y - X\beta)^2$, again differentiating w.r.t. β we get:
  $-X(Y - X\beta) = 0 \Rightarrow X^T Y = X^T X\beta$

- So, again our model is $Y = X\beta + \epsilon,\ Y \sim \mathcal{N}(\mu, \mathbf{I}\sigma^2), \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$
  For OLS estimate we have to minimize $(Y - X\beta)^2$, again differentiating w.r.t. β we get:
  $-X(Y - X\beta) = 0 \Rightarrow X^T Y = X^T X\beta \Rightarrow (X^T X)^{-1} X^T Y = \beta$
- The quantity $(X^T X)^{-1} X^T$ here is called *Moore Penrose Pseudoinverse* and can be thought of as equivalent of an inverse matrix for non-square matrices.

References I

[1] Bruno De Finetti. *Theory of probability: A critical introductory treatment*. Vol. 1. John Wiley & Sons, 1970.

[2] N. Reid. "Likelihood". In: *Journal of the American Statistical Association* 95.452 (2000), pp. 1335–1340.