Energy Analytics

Adam Carrera

January 20, 2021

CONTENTS

Part I

Week 1

LECTURE

1 COURSE INTRODUCTION

1.1 Syllabus

The course syllabus is available on eLearning and coursebook. Professor and TA office hours are in MSTeams. Every Tuesday and Thursday 10 am - 11 am and 3 pm - 4 pm. Lectures will be posted Mondays and Wednesdays on MS Stream.

1.1.1 Homework

Homeworks are coding assingments that are submitted to elearning. Schedule is available in the syllabus.

1.1.2 Exams

The final exam will be open book. It will be assigned and submitted thorugh elearning.

1.1.3 Final Project

Recorded presentation and report through MS Teams and eLearning.

1.1.4 Course Description

Understand the basic concept of analytics, including descripting analytics and predictive analytics. Understand applications of analytics in energy industry. Know how to use mathematical software and other commercial packages. Know how to develop load, price, wind and solar forecasts.

1.2 Discussion

1.2.1 Who makes up the most emissions?

China takes up the biggest share of global emissions at about 30%. Electricity and heat make up 24.9% of world greenhouse gas emissions (highest portion). There are several end uses for electricity and heat like buildings, refining, food. These processes create lots of CO_2. The majority of CO2 emissions in the US come from the Electricity sector, followed by transportation, and industry.

1.2.2 How do renewable technologies compare?

Renewable and carbon neutral technologies are utilized a lot less than coal and natural gas. Coal use has gone down from about 50% of electricity generation in the US to 32.3% (less than one third!). Natural gas however, has increased from 21.4% to 32.3%. Nuclear energy has remained constant throughout the years, due to the effrot required to build the plant. Hydropower has also remained constant through out the years. Other renewable sources are growing very fast (solar, wind power, biofuels, etc).

1.2.3 Wind and Solar Energy

Wind has the higest capactity of non-hydro renewables. By the end of 2018 the world was generating 591 Gigawatts of wind energy. Wind energy has also been growing very fast in the US as well. In Q3 2019 the US generated 100,128 MW, with Texas leading at 27,000 MW. However in some parts of the US, the conditions are not suitable for wind turbines.
PV (solar energy storage) has been steadily growing in the US as well. PV installations are split between Residential, Non-residential, and Utility. Solar energy resources vary with location as well.

1.2.4 What are the biggest drivers of renewable installations and penetration rates?

Policy is a big driver, states have different renewable portfolio standard targets. The costs to implement renewable energy has been decreasing over the years. Solar is a little more expensive than wind, but they are both becoming competitive with non-renewables. Utility-Scale PV costs have droped from $5.52 per W to $1.13 per W.

1.3 Renewables Integration

Power generation and energy consumption must be balanced. This balance becomes more challenging with renewables, because you cannot control how much wind blows or how much the sun shines.

Part II

Week 2

LECTURE

2

ENERGY ANALYTICS AND R

2.1 Energy Quiz

1. How much of the energy in burning coal reaches the consumer as electricity? (33%)
2. Which State consumes the most energy? (Texas)
3. Which state produces the most coal? (Wyoming)
4. What country produces the most coal? (China)
5. Which country generates the most electricity from nuclear power? (United States)
6. What country generates the greatest share of its electricity from wind power? (Denmark)
7. Electricity is called a secondary energy source because ... (We get it from converting other energy sources)
8. Most of the energy we use originally came from? (The sun)

2.2 Power System Background

Power systems have transitioned from centralized to distributed generation (Solar panels on roofs, batteries, PV panels, small generators). One challenge is that the energy flows in two directions, instead of one. Another challenge is that renewable energy is variable and uncertain. However, advanced communication has allowed decentralized control of devices.

2.2.1 Power System Operations Timescales

The power system load changes throughout the day. As your load changes, your power generation should change as well. Regulation is seconds to minutes, load flowing is minutes to hours, and scheduling is the entire day. Unit commitment refers to several day.

2.2.2 Energy Sector

Electricity is only one part of the energy sector. There is a trade off between fixed costs and operating costs.

- Nuclear and Solar have high fixed cost, low operating costs
- Natural gas/oil have low fixed costs, high operating costs (depends on fuel prices)
- Coal, wind, hydro are in between

A units **Capacity Factor** is important to determining ultimate cost of electricity. Wind has a low capacity factor (50%). Some renewables have tax credits, which affect the economics. Different renewables have different levelized costs (USD/kwH). The levelized cost of Solar PV has decreased from 2010 to 2014.

2.2.3 Conventional Power System

The conventional power system is a one way system. You generate the power and send it to the customer. In later lectures we will talk about different changes in power systems to centralized systems to decentralized systems.

1. Generation (Source)
2. Transmission Substation (Change voltage to a suitable value)
3. Transmission System (Carry energy long distances)
4. Distribution Substation (Change volatge, smaller than transmission substation)
5. Distribution System (Power lines, distribute to customer)
6. Customer (Loads)

2.2.4 Power Plant Classification

There are 3 categories.

1. Base-Load Generating Units
    - Constant power generation over time
    - Long start-up and shut-down times
    - Low cost of production, but high capital investment
    - ex. Nuclear, Hydroelectric, High-Performance Steam Turbine Plant
2. Intermediate (Load Following) Generating Plant
    - Typicall runs at full-load between 2,000 - 5,000 hrs/yr
    - Operation and economic characteristics somewhere betwwen base-load and peak-load
    - ex. Simple Steam Turbine Plant, Old Base-Load Plant, combined Gase and Steam Turbine Plant, Sleected Hydroelectric Power Plant
3. Peak-Load Generating Units
    - Quick starting and shutdown capability
    - Runs between 100 - 2,500 hrs/yr during peak loads
    - Standby and emergency units
    - ex. Gas Turbine, Diesel Engine, Hydro - Pumped Storage

2.2.5 Interconnections

The US has 3 synchronous grids (interconnections). The Eastern Interconnection, Western Interconnnection, Texas Interconnection (cool!).

2.2.6 Renewable Energy and Distributed Generation

Most renewable energy will be converted into electricity. The resource will be distributed geographically and mostly dependent on changing weather and climate. However, the resource cannot be directly controlled in the manner of conventional generation (variable and uncertain output).
Renewable resources are more distributed than conventional types. They are often connected at low voltages. Increasing the input from renewable energy sources requires a revision of the way power systems are designed and operated in order to better accomodate these sources better.

2.3 Analysis

2.3.1 Actionable Decision Support & Automation

The amount of analytics and human input depends on the type of analysis. Our goal is to use prescriptive analytics, so the computer will either support a decision, or make the decision automatically.

- Descriptive - What happened?
- Diagnostic - Why did it happen?
- Predictive - What will happen?
- Prescriptive - What should I do?

2.3.2 Behind the Meter (BTM) Solar Forecasting

Different levels of difficulty associated with solar forecasting. BTM is for very small systems.

- Bottom-up approach
- Multi-timescale Forecasting
- Distributed Solar Ramp forecasting

2.3.3 Types of Load Forecasting

- Short term forecasts (one hour to a week)
- Medium forecasts (a month up to a year)
- Long term forecasts (over one year)

2.3.4 Key Methods

- Time series models, visualization, regression models, decomposition
- Machine Learning models
- Deep learning models
- Probablility, distribution, visualization

2.4 Distributions

We will talk about probability distributions more in-depth in the future.

2.4.1 Normal Distribution

- Probability density function
- Cumulative distribution function

LECTURE

3

R IN CLASS

Part III

Week 3

LECTURE

4

CONVENTIONAL POWER GENERATION

4.1 Thermal Power Generation

- Coal-fired thermal power: Baseload operation
- LNG-fired thermal power: Base to middle load operation (Liquid Natural Gas)
- Oil-fired thermal power: Middle to peak load operation

The basic principle is,

- Fuel tank sends fuel to boiler
- Boiler boils water which creates team
- Turbine is driven by the high temperature steam
- Generator is powered by the turbine

4.1.1 Coal Generation Details

Coal plants are divided into Regular, Super Critical, and Ultra Super Critical plants. The pressure levels, efficiency levels, and temperature increase with each plant.

Plant Name	Energy Source
Grand Coulee	Hydroelectric
Palo Verde	Nuclear
Martin	Natural gas, fuel oil

Table 4.1: Largest US Power Plants and Primary Energy Source

4.1.2 Grid Integration Details

Coal plants are base load, which means long start up and shutdown times (24 hrs or more!). The long minimum run times and high minimum generation levels are a challenge for integration.

4.1.3 Plant Scherer - Georgia

- Total Capacity: 3,520 MW
- Largest GHG emitter in the US (20M tons per year)
- 7th largest plant in the US (by Capacity)
- Capacity Factor: 61.2 %
- Consumes 11M tons of coal per year
- Each of four cooling towers circulates 270k gallons of water per minutes with 8k gallons lost to evaporation

4.2 Discussion

- What are the three largest generation plants in the US?
- What technologies would you expect them to be?

See Table 4.1.

4.3 Natrual Gas Plants

Conversion efficiencies typically between 20 - 35% with temperatures of 2000 F. These turbines can generally start up and shut down faster, as well as being able to ramp up more quickly. May be flexible with fuel oil as an alternative.

4.3.1 Turbine Types

- Heavy Frame
  - Lower pressure ratio (< 20)
  - Tend to be larger
- Aero-derivative
  - Higher pressure ratio (> 30)
  - Tend to be smaller

4.4 Discussion

Gas and steam turbines have low efficiencies. How can we increase that efficiency?

- We can combine the gas and steam turbines, this is called a combined cycle.
- Another way is to have the gas from the gas plant boil the water for the steam plant.

The efficiency can be increase to about 50 - 60%. Can be operated seperately as well. The typical start-up time for (older) full units is 2-4 hours. These are historically mid range units.

4.5 Nuclear Generators

Nuclear is a carbon neutral technology.

- Containment building holds the reactor, which boils the water
- The steam powers the turbine, generating electricity

4.5.1 Nuclear Facts

- Over 400 plants in 30 countries
- Power form Fission, not fusion
- Uses uranium 235
- Nuclear waste disposal is still an unsolved problem

4.5.2 Grid Integration

- Base load, week long startup and shutdown times
- Very little ramping capability
- Large capacity plants
- Ultra-high minimum generation levels, often 90% capacity

4.6 Hydropower

- Conventional - dams
- Unconventional hydropower
  - Hydrokinetic power

Which country produced the most hydropower?

1. China
2. Brazil
3. Canada
4. US
5. Russia

We are unlikely to build more conventional hydro any more (environment concerns). But we can still build pumped storage. Here are some more facts.

- 80 - 90% efficiency
- 6 - 10 % of US electricity generation
- Roughly 100 GW of capacity
- 80k dams in the US, 2.4k have hydropower
- about 20% of world electricity usage
- 77% of Brazilian power
- about 17% of Chinese power
- 95% of power in Norway

4.7 Discussion

What are some of the limitations towards expanding hydropower in the US? Internationally?

4.8 Geothermal Generation

Cold water is pumped down to the earth. The heat of the earth heats up the water making steam, which travels up to the generating station.

- Dry steam (150 C)
- Flas steam (180 C)
- Binary Cycle (60 C)

LECTURE

5

PROBABILITY & DISTRIBUTION

5.1 Probability Density Function (PDF)

Suppose we have some variable $X \sim f(x)$ where $f(x)$ is the probability density function (pdf) of X.
The requirements on $f(x)$ are:

- $f(x) \geq 0 \quad \forall x \in X$. Where X is the domain of X.
- $\int f(x)dx = 1$.

The normal distribution pdf has the form,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}.$$

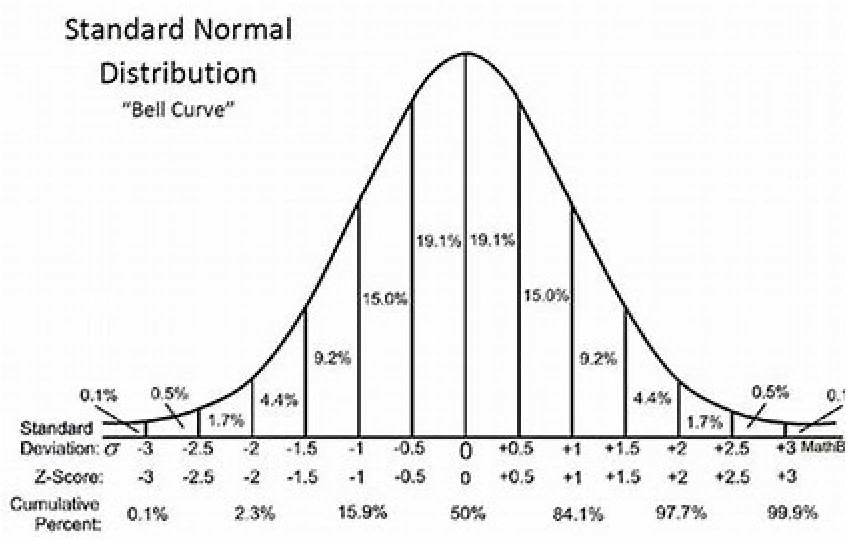

Figure 5.1: Standard Normal Distribution

5.2 Standard Normal Distribution

If $X \sim N(0,1)$, then X follows a normal distribution.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{5.1}$$

5.3 Cumulative Distribution Function (CDF)

The CDF $F(x)$ for a discrete random variable X gives, for any specified number x the probability $P(X \leq x)$. Obtained by summing the pdf $p(y)$ over all possible values y s.t $y \leq x$. The CDF of a continuous random variable gives the same probabilities $P(X \leq x)$ and is obtained by integrating the pdf $f(y)$ between minus infinity and x.

$$F(x) = P(X \leq x) = \int_{\infty}^{\infty} f(y)dy \tag{5.2}$$

For each x, F(x) is the area under the density curve to the left of x.

5.3.1 Obtaining PDF from CDF

If X is a continuous random variable with pdf $f(x)$ and cdf $F(x)$, then at every x at which the derivative $F'(x)$ exists,

$$F'(x) = f(x).$$

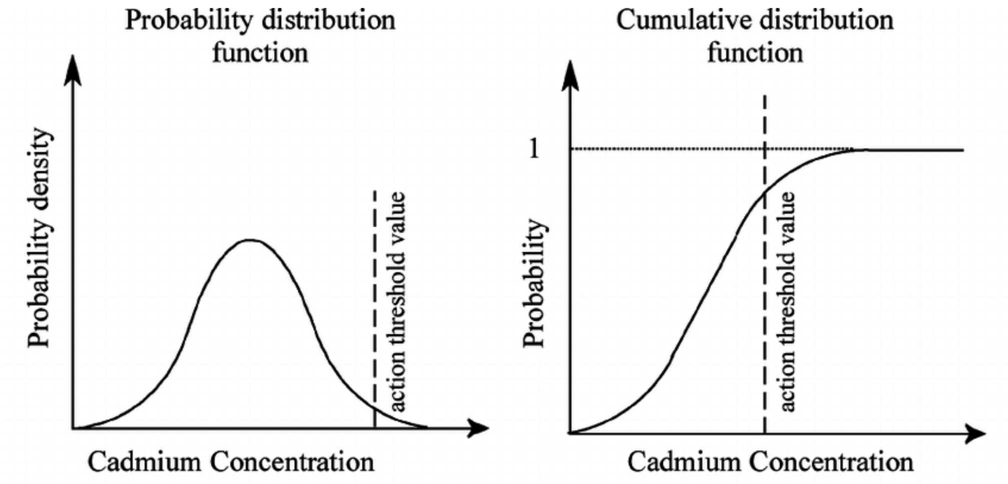

Figure 5.2: A PDF and associated CDF

Table 5.1: Properties of Normal Distribution

PDF	see above section
CDF	$\frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right).$
Quantile	$\mu + \sigma\sqrt{2}erf^{-}1(2F-1).$
Mean	$\mu.$
Median	$\mu.$
Mode	$\mu.$
Variance	$\sigma^2.$

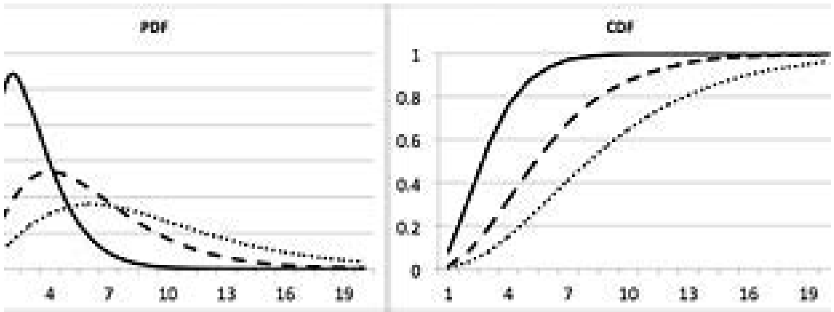

Figure 5.3: Gamma PDF and CDF

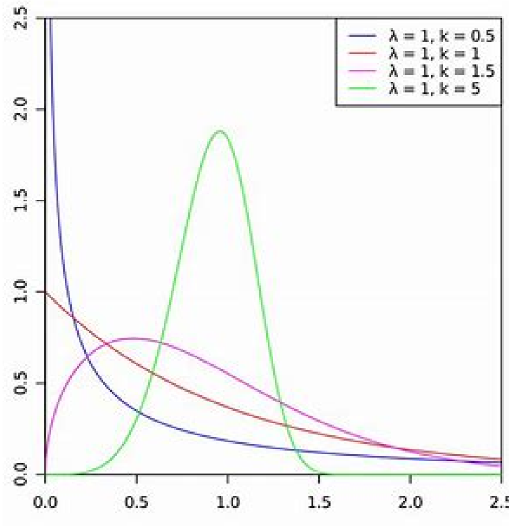

Figure 5.4: Weibull PDF

5.4 Gamma Distribution

PDF:

$$f(x;k,\theta)=\frac{x^{k-1}e^{-\dfrac{x}{\theta}}}{\theta^k\Gamma(k)},\quad x,k,\theta>0 \tag{5.3}$$

CDF:

$$F(x;k,\theta)=\int_0^x f(u;k,\theta)du=\frac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)} \tag{5.4}$$

where k is the shape parameter, and θ is the scale parameter.

5.5 Weibull Distribution

5.6 Problem of Estimation

We want to estimate $f(x)$ or $F(x)$ from a smaple of data $\{x_j\}_{j=1}^{n}$.
Different approaches

1. Histogram estimates
2. Kernel density estimates
3. Single distribution estimates

5.7 Single Distribution Estimates

Use R package: fitdistrplus. KDE with Gaussian kernel

Part IV

Week 4

LECTURE

6

WIND ENERGY ANALYTICS

6.1 Wind Map Information

- Wind map is digitized using image processing tools
- Total area under different average wind speeds (AWS) is estimated with an interval of 0.5 m/s
- A normal distribution is fitted to represent the geographical distribution of average wind speeds over contiguous USA

$$\mu = 5.6 \text{ m/s}, \quad \sigma = 1.3 \text{ m/s}.$$

6.2 Discussion

What are the five largest wind turbine manufacturers in terms of global market share?

- Goldwind
- Vestas
- GE Wind
- Siemens

- Gamesa

6.3 Wind Generator Types

- Type 1 - Fixed-speed wind turbine
- Type 2 - Variable-slip wind turbine
- Type 3 - Pitch controlled wind rotor
- Type 4 - Full converter wind turbine

Type 1 has a squirell cage induction generator, type two has a wound rotor induction generator and a control signal.

6.4 Offshore Pros and Cons

Advantages

- Generally higher capacity factors
- Less visual impact
- Often closer to load centers
- Better correlation with load in some locations

Disadvantages

- Higher installation costs
- Maintenance issues

Part V

Week 5

LECTURE

7

TURBULENCE

7.1 Wind Energy - Turbulence

Wake loss leads to significant energy loss, about 5 to 20 percent from the whole plant. Turbulence is characterized by ambient and wake turbulence.

- ambient turbulence - normal turbulence at the site that weould be experienced by one turbine
- wake turublence - caused by upwind turbines shading those downstream

How do the turbulence conditions affect wind power differently?

7.1.1 Monitoring Data

Using Xcel Cedar Creek wind plant data, we can

- determine data sets for in and out of wake
- construct surrogate models $P = f(U), P = f(U, TI)$. (TI - Turbulence Intensity)
- Quantify the uncertainty in turbine power generation

7.1.2 In and Out of Wake Scenarios

Pick one turbine and calculate which turbines are in its wake based on wind direction.

7.1.3 Surrogate Modeling: In-Wake and Out-of-Wake Scenarios

For each set of data, two surrogate models were developed to represent the turbine power generation. As a function of wind speed and as a function of both wind speed and turbulence intensity.
The turbulence intensity is defined as the standard deviation of the wind speed within a short time period divided by the mean wind speed during that time period.

$$TI = \frac{\sigma_U}{U} \tag{7.1}$$

7.1.4 Support Vector Regression

Surrogate modeling is concerned with the construction of approximation models to estimate system performance and to develop relationships between specific system inputs and outputs.

7.1.5 Uncertainty in the Power Response

Once we have our models we can quantify uncertainty.
The discrepancy between the actual power generation and the power curve estimation of a wind turbine can be attributed to the following major factors:

- Wind shear - the variation of wind with vertical distance from the ground
- Turbulence effects - the power generation depends on both mean wind speed and turbulence
- Turbine reliability - the uncertainty in the turbine performance

We want to illustrate how the uncertainty is different between in and out of wake scenarios. This is measured by the uncertainty with respect to the reported power curve

$$E_{pc} = P_c(U) - P_g(U) \tag{7.2}$$

Where Pc is expected power and Pg is recorded power. Epc is the error of power curve. If we wanted to estimate based on only wind speed we would have. The uncertainty in power generation with respect to estimated power response that only accounts for mean wind speed. **These are both error distributions.**

$$E_{pf} = P_g(U) - P_f(U) \tag{7.3}$$

Where Pf is the surrogate estimation. Finally, we

LECTURE

8

REGRESSION

For the homework, only one year of site B data has been given. You will need to compare the accuracy of your MCP methods based on the data. You can use 11 months for correlation and the rest of the months for validation. Review MCP slides on training and validation.
The book has some helpful material! Chapter 5: Time series regression methods.

8.1 Correlation

Correlation coefficients are used to measure the strength of the relationship between two variables.

$$r = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2(y_t - \bar{y})^2}} \tag{8.1}$$

xt is the value at the time, x bar is the mean value of the data. R will lie in between -1 and 1, where a value of 1 indicates perfect positive correlation. Note that correlation values can have many different visual representations. Y should be a function of some variable that has a higher correlation coefficient with the target variable

8.1.1 Autocorrelation

Autocorrelation measures the linear relationship between lagged values of a time series. The correlation coefficient is calculated for different amounts of time delay. This is important for forcasting future events. If the historic data has a high correlation with the future event, it will be easier for you to make an accurate forcast.

8.2 Time Series Regression Models

Basically, we forecast the time series of interest y, assuming that it has a linear relationship with other time series x. The forecsat variable y is sometimes also called the regressand, dependent, or explained variable. The predictor variables x are someteiems called the regressors, independent or explanatory variables.

8.2.1 Linear Model

Simplest case is the linear relationship.

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \tag{8.2}$$

β_0, β_1 are the intercept and slope of the line. Epsilon is the error.
Note,

$$\beta_0 + \beta_1 x_t \equiv \hat{y}.$$

Plot the points, draw the line!

8.2.2 Multiple linear regression

Two or more predictor variables.

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_k x_{k,t} + \epsilon_t \tag{8.3}$$

Each predictor varaible must be a numerical value. Each coefficient measures the effect of each predictor after taking into account the effects of all the other predictors in the model. Thus, the coefficients measrue the marginal effects of the predictor variables.

8.3 Data Fitting

Data fitting is the process of constructing a mathematical function to fit to a series of data points. Data poitns can be experiment results, simulation results, or observations of phenomena. There are many methods for dta fitting. The least-squares appraoch, which requires the use of optimization is introduced here.

8.3.1 Least squares estimation

We need to estimate β coefficients. Least squares helps us choose those by minimizing the sum of the squared errors. Lots of machine learning problems use this.

$$\min \sum_{t=1}^{T} \epsilon_t^2 = \min \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_{1,t} - \beta_2 x_{2,t} - \ldots - \beta_k x_{k,t})^2 \tag{8.4}$$

8.3.2 Mathematical Example

In most cases we will not be able to find all of our unknowns to solve the system of equations. Can even have millions of design variables. Least squares can be used to solve these types of problems! The curve will be very close to all of the points, but will not have all of them.

8.4 Goodness of fit

To test how well the linear regression model fits, we find the coefficient of determination.

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2} \tag{8.5}$$

R^2 will lie in between 0 and 1. If the predictions are close to the actual values, we would expect R^2 to be close to 1.

8.5 Nonlinear Regression

Linear regression is used alot, but in practice most problems are non-linear. The simplest way of modeling a nonlinear relationship is to transform the forecast variable y and/or the predictor variable x before estimating a regression model. While this provides a non-linear functional form, the model is still linear in the parameters. THe most commonly used transformation is the natural logarithm.
A log-log functional form is specified as,

$$\log y = \beta_0 + \beta_1 \log x + \epsilon \tag{8.6}$$

8.6 Support Vector Regression

Surrogate modeling is concerned with the construction of approximation models to estimate the system performance and to develop relationships between specific inputs and outputs.
SVR is a regression method to construct smooth, nonlinear, regression approximations by formulating the surrogate model construction probelm as a quadratic programming problem.

8.6.1 Radial Basis Function

Expresses surrogate models as linear combinations of aparticular type of bassi function $\psi(r)$. Each constituent bassi function is defined in terms of the Euclidean distance (r) between a training point and the poitn of evaluation. (r) can be expressed as

$$r = ||x - x_i|| \tag{8.7}$$

x is the point of evaluation, x_i is the ith training point. The RBF model is then expressed as a linear combination of the basis functions across all training points.

$$\tilde{h}(x) = \sum_{i=1}^{m} \omega_i \psi\left(||x - x_i||\right) \tag{8.8}$$

ω_i are the generic weights of the basis functions which are determined by training. Ψ is a matrix of the basis function values at the training poitns. and W is a vector containing the weights. Y is a vector of the function values at the training points. The weights ω_i are found by solving.

$$\Psi W = Y \tag{8.9}$$

RBFs are essentially interpolating functions. A trained RBF model will pass through all the training points.