GRADUATE SCHOOL
OF ARTS AND SCIENCES
COLUMBIA UNIVERSITY

Machine Learning scratching

Linear Regression guidebook

Name: Zhiyi Chen

Email: zc2489@columbia.edu

Date: January 24th 2021

Contents

1 Linear Regression Basic Concept

1.1 Random Variable, Parameter and Model

As a preparation to dive into linear regression, there are three basic terminologies highly necessary to be fully understood, that is, random variable, parameter and model.

Random Variable. According to Wikipedia[10], in probability theory context, a random variable is understood as a measurable function defined on a probability space that maps from the sample space to the real numbers. From this description, there are two points concentrated that

- Random variable is a map function, the input is sample space ω, a random variable X can be written as $X(\omega)$
- The value of random variable is unknown[7] since it represents an outcome of an event at one time, at another time the value may be different which we never know. Here comes the **randomness**. Sometimes we may encounter a formula that $X_1 = 20$, does it mean X_1 is not a random variable anymore since it obtains value 20? of course not, it means this time people capture a observation valued 20 from sample space, next time the captured value cannot be certain. However we can annotate 20 with low-case x_1 defined as a data choosing.

The definition of random variable can be shown in following graph[10]:

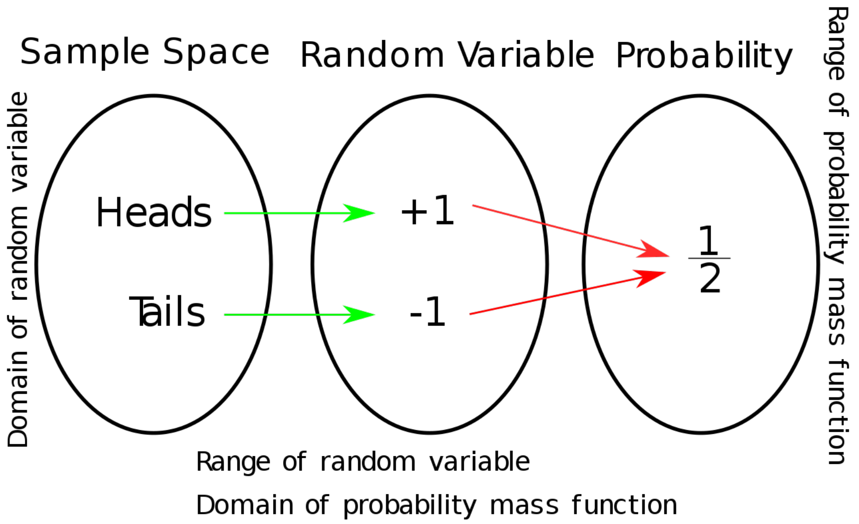

Sample Space
Random Variable
Probability
Domain of random variable
Heads
Tails
+1
-1
1/2
Range of probability mass function
Range of random variable
Domain of probability mass function

Figure 1: Definition of random variable

Parameter. Parameter is a bridge linking random variable and model. In statistics, distribution of random variable can be described by parameters. Model applies parameters to form its structure. The estimation of model parameter should be a statistic composed with data and information. Sometimes we may hear a word called hyperparameter, the division between it and parameter is not that strict. Model parameters are estimated from data automatically and model hyperparameters are set manually and are used in processes to help estimate model parameters[1].

Model is the core research target in linear regression, here we use a simple linear regression to explain how it works and how does random variable and parameter link to model.

First we should think why we need a model, in another way, what's the goal of building a model. The motivation of constructing model is to find a pattern among all target data, i.e. population which we never know. For example, we have a research on the relationship about the house price in Shanghai and the distance between house and Shanghai city center. Here, population is the prices of all houses in Shanghai and we assume the relation is a simple regression. Normally there will be two model structures displayed.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, ..., n \tag{1}$$

$$\epsilon_i \sim^{iid} N(0, \sigma^2)$$

where i we can understand it as the order of data choosing. x_i, not a random variable, means a value of distance at i-th choosing, ϵ_i is a independent and identically distributed random variable which means the error term at i-th choosing. As for random variable, there is an useful property that

- If X is a random variable, g(X) is also a random variable where g() is a function.

Therefore, Y_i is also a random variable. This model structure does not reflect all data relevant to our topic, it shows one layer of population data. By combining our assumption background above, it can be explained as when the distance between house and center of Shanghai is equal to x_i, the price of house can be expressed as a simple regression response. To explain all data, the model structure should be

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2}$$

$$\epsilon \sim^{iid} N(0, \sigma^2)$$

where we remove all i and replace x as capital one X. In order to express all data, we have freedom to choose all possible distance, not just one value, thus our predictor should be a random variable. It is a generalization model structure and we can just understand it as the price of house in Shanghai is a simple regression expression of distance away from city center. Easy and visulization explanation is shown blow:

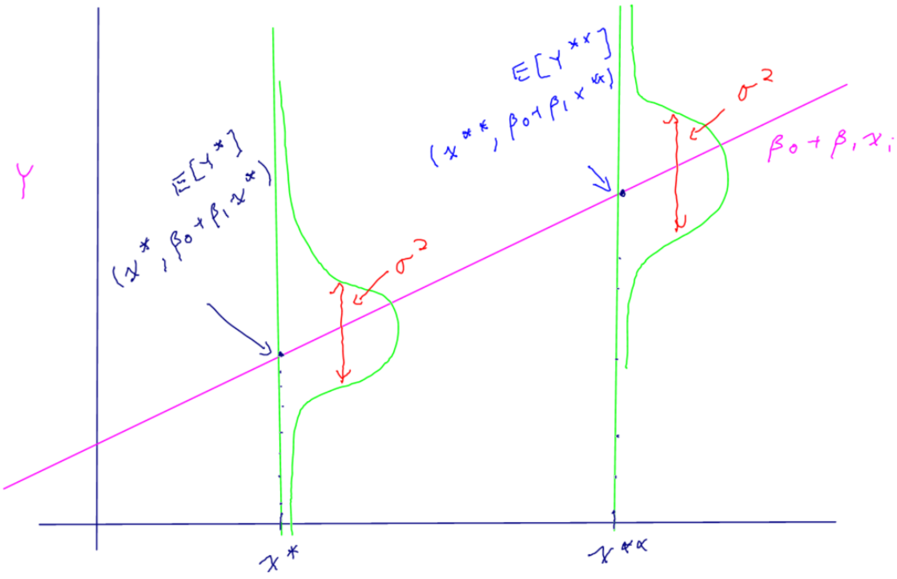


Figure 2: Simple linear regression explanation

However, since we never know all data in population, β_0 and β_1 will always be unknown parameters. So, the question comes that now that population and parameter are mysteries, how do we build the model and what's the meaning of it. Here comes the sample data, all datasets we can see from books and any other materials are all sample data, from which we observe their distributions, make assumptions, build proper models, evaluate performance and contribute improvement. Therefore, our study method is that

- **Population will mimic sample under certain assumptions.**

1.2 Degree of Freedom

By the definition given by Wikipedia, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary[3]. For beginners, it is not easy to understand since "free to vary" is a very confusing term. Here we will give a simple example to explain by step.

Situation 1. There are 10 samples X_1, X_2, ..., X_{10} from the same population, the mean of them is 20. What is the number of degree of freedom in the mean $\overline{X} = \frac{X_1+X_2+...+X_{10}}{10}$[4].

First confusing point is that we already know the mean value, is it possible to have something free to vary? We should pay attention that all samples are random variables with unknown value. Even though we obtain the mean value which means each variable takes one value at this time, we should concentrate on the variable equation since variables may take different value at other times.

$$\frac{X_1 + X_2 + ... + X_{10}}{10} = 20 \tag{3}$$

To make it accurate, variables need to take values, thus only 9 variables have freedom to choose and the last one value can be calculated as a fixed number.

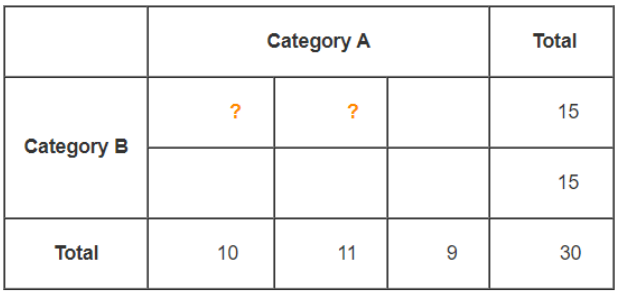
Situation 2. There are n samples X_1, X_2, ..., X_n from the same population. What is the number of degree of freedom in the mean $\overline{X} = \frac{X_1+X_2+...+X_{10}}{10}$

The only difference between situation 2 and situation 1 is that it does not tell the mean value, but it does not matter since the mean $\overline{X}$ must take one value, then the logic is the same as previous one.

Situation 3. There are n samples X_1, X_2, ..., X_n from the same population with a mean $\overline{X}$. Define a statistic $S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n-1}$, calculate its degree of freedom number.

Since $\overline{X} = \frac{X_1+X_2+...+X_n}{n}$, it is no doubt that S_X^2 is also a combination of $\{X_i\}_{i=1}^{n}$, so the situation is the same as previous one, the number of degree of freedom is n-1.

Situation 4. A chi-square test of independence is used to determine whether two categorical variables are dependent. According to the table shown below, it is easy to judge visually that if we know any two values among fills, we can know all others value. Thus the degree of freedom number is (2-1)(3-1) = 2.



	Category A			Total
Category B	?	?		15
				15
Total	10	11	9	30

Figure 3: Situation 4 example

Situation 5. There are n samples (x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n) from the same population. They take y_1, y_2, ..., y_n as value respectively. After observing the distribution of these values, we decide to apply simple linear regression to express the pattern $\hat{Y}_i = \beta_0 + \beta_1 x_i$. According to least square method, we obtain "best" parameter $\hat{\beta}_0$ and $\hat{\beta}_1$. What is the number of degree of freedom of MSE $= \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$.

First we need to know the expression of "best" parameters

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

After calculation, we can know the value of $\hat{\beta}_1$ and $\hat{\beta}_0$. Then it comes to a linear equations with two unknowns problem. If we know n-2 values of random variables, then value of rest two are fixed by solving the equations above. The number of degree of freedom is n-2.

Conclusion. Degree of freedom are through all statistic process, we should have a very clear understanding upon it.

- For situation 1-4, they are all about non-parametric statistic. Assume a non-parametric statistic has n dimensions, the number of dimensions are $n_1, n_2, ..., n_n$ respectively, then the number of degree of freedom is $\prod_{i=1}^{n}(n_i - 1)$.

- For situation 5, it is about parametric statistic. Assume there are n variables and m parameters, the number of degree of freedom is n-m.

1.3 Hypothesis tests

From the knowledge learned from school, hypothesis test seems to be not that difficult since we can just follow the steps as an instruction to give a conclusion. However, what's the meaning of confidence interval, what influences confidence level, why it also always be 95%, how should we explain practical problem by hypothesis test and which test is proper to use during hypothesis test. Those question are all important in linear regression applications. In this part, we will use examples to explain clearly how hypothesis test works and finally we can discover that in our daily life, it is applied everywhere.

Situation. According to *Status of Nutrition and chronic Diseases among Chinese residents report (2020)*, the average height of Chinese male aged 18-44 is 169.7cm, of female during this age range is 158cm.

Does it mean the truth? Maybe you will think what about the people whose height information are not recorded. Of course not, 169.7cm and 158cm are just one estimate, a value from sample. Just in case we need to recognize the difference between estimate, estimator and estimation. Estimator is an expression of random variables. Least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are examples. Estimate is one possible value of estimator, normally, we can calculate it by obtaining values of random variables. Estimation is the process from estimator to estimate.

Let's arrange knowledge linking to this situation and use male condition for example.

- Height of Chinese male aged 18-44 is **Population** since it is impractical to gather all height information about Chinese male aged 18-44.
- The average height of Chinese male aged 18-44 is an **Parameter** μ, which we never known.
- Assume sample size is n. Random variable $X_1, X_2, ..., X_n$ means the height of Chinese male aged 18-44. The estimator of μ is $\bar{X}$.
- At this sampling, we align $x_1, x_2, ..., x_n$ to these random variables according to a certain sampling method and use the mean of $x_1, x_2, ..., x_n$ as an estimate $\bar{x}$ which is equal to 169.7cm.

Why we can use 169.7cm to represent parameter μ, since it is within confidence interval or not rejected by hypothesis test.

1.3.1 Confidence Interval

Situation continue 1. Under 95% confidence level, the confidence interval of parameter μ is [166.7,172.7].

One sampling gives one estimate, but we cannot say only this one value can represent μ since $\bar{X}$ is an estimator, we should consider its standard error to give an interval. It means that the we cannot reject the value within the interval to represent true mean μ under certain **confidence level**.

Confidence level is a significant idea in statistics fully considering the difference between population and sample. Take the 95% confidence level mentioned in situation as an example. If we have 100 samplings which gives 100 confidence intervals, there will be 95 intervals containing the true parameter. However, we do not know which intervals contain the true parameter. Thus, we can understand it in another way that each confidence interval has 95% probability to contain the true parameter. It does not mean the true parameter must be in the interval, but we, investigators have 95% confidence to obtain the true parameter. So for a certain estimate, which values shall we put 95% "believe" on? The values having top 95% probability in estimator's distribution.

Situation continue 2. Assume $X_1, X_2, ..., X_n \sim^{iid} N(\mu, \sigma^2)$ where μ is unknown, σ^2 is known. How to calculation the confidence interval of μ under 95% confidence level. According to the explanation above, what we need is the shaded interval. Due to the distribution of $\{X_i\}_{i=1}^n$, we can know the distribution of $\bar{X}$ is $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

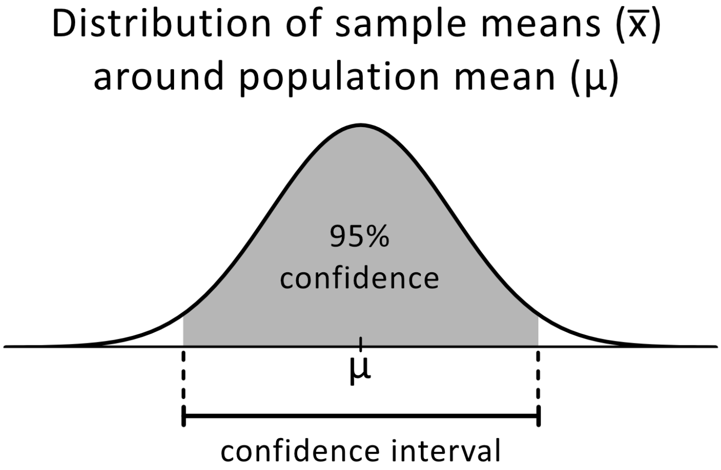


Figure 4: Confidence Interval based on 95% confidence level

Assume the lower bound of $\bar{X}$ is l, upper bound of $\bar{X}$ is u. Then we can build equations that

$$P(\bar{X} \leq l) = 0.025 \Rightarrow P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{l - \mu}{\sigma/\sqrt{n}}) = 0.025$$

$$\frac{l - \mu}{\sigma/\sqrt{n}} = z_{0.025} \Rightarrow l = \mu + z_{0.025}\sigma/\sqrt{n} = \mu - z_{0.975}\sigma/\sqrt{n}$$

The same solution for upper bound,

$$P(\bar{X} \leq u) = 0.975 \Rightarrow P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{u - \mu}{\sigma/\sqrt{n}}) = 0.975$$

$$\frac{u - \mu}{\sigma/\sqrt{n}} = z_{0.975} \Rightarrow u = \mu + z_{0.975}\sigma/\sqrt{n}$$

Then for $\bar{X}$, the confidence interval is $[\mu - z_{0.975}\sigma/\sqrt{n}, \mu + z_{0.975}\sigma/\sqrt{n}]$, after simple transformation, the confidence interval of μ is $[\bar{X} - z_{0.975}\sigma/\sqrt{n}, \bar{X} + z_{0.975}\sigma/\sqrt{n}]$. After one sampling, we can obtain interval $[\bar{x} - z_{0.975}\sigma/\sqrt{n}, \bar{x} + z_{0.975}\sigma/\sqrt{n}]$.

What will affect Confidence Interval.

- Sample size. As mentioned before, confidence level reflects the difference between population and sample. The smaller your sample, the less likely it is you can be confident the results reflect the true population parameter or sometimes we may need to change our test method, such as t test or others.
- Confidence level. Actually it is highly linked to sample size, less confidence level should be given when sample size goes to small one. Here we want to mention 0% and 100% confidence level. 0% confidence level happens when investigator have no faith at the sampling, that is, our sampling is meaningless. 100% confidence level comes true only when sample is equal to population. In this dynamic world also containing rules and patterns, these two conditions are not within our consideration.

1.3.2 z test and Student's t test

Situation original form. Assume $X_1, X_2, ..., X_n$ are iid taken from population in normal distribution with mean μ and variance σ^2. $\bar{X}$ is the sample mean. At one sample, we obtain mean $\bar{x}$ and standard deviation s. We want to do hypothesis test that

- $H_0 : \mu = \mu_1$
- $H_1 : \mu \neq \mu_1$

We will skip the process of hypothesis test since it is easy to be understood and the core is test method. Here we want to emphasis the z test and student's test, compare their application in real life.

- z test: $z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
- t test: $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$

The common assumptions of these two test are

- The data is collected from a representative, randomly selected portion of the total population.
- The data, when plotted, results in a normal distribution, bell-shaped distribution curve.

z test asks for a known population variance and sufficiently large sample size. However, t test doesn't ask for population variance and the sample size doesn't need to be very large as long as investigators is able to recognize the shape of sample is bell-curved. Therefore, t test should be more practical than z test since population information is unknown.

Situation central limit theorem. Assume $X_1, X_2, ..., X_n$ are iid taken from population with mean μ and variance σ^2. $\bar{X}$ is the sample mean. At one sample, we obtain mean $\bar{x}$ and standard deviation s. We want to do hypothesis test that

- $H_0 : \mu = \mu_1$
- $H_1 : \mu \neq \mu_1$

The only difference between this situation and previous one is that we don't specify the distribution type of population. Here according to CLT, $\bar{X}$ is $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. CLT only works for $\bar{X}$.

2 Linear Regression procedure

2.1 Data Analysis guideline

Linear regression is for data analysis, from my experience, a brief procedure for data analysis can be described in following map. If we choose linear regression as the main character in our study or research topic, how should we go through the map. In this part, more details of my analysis idea will be explained.

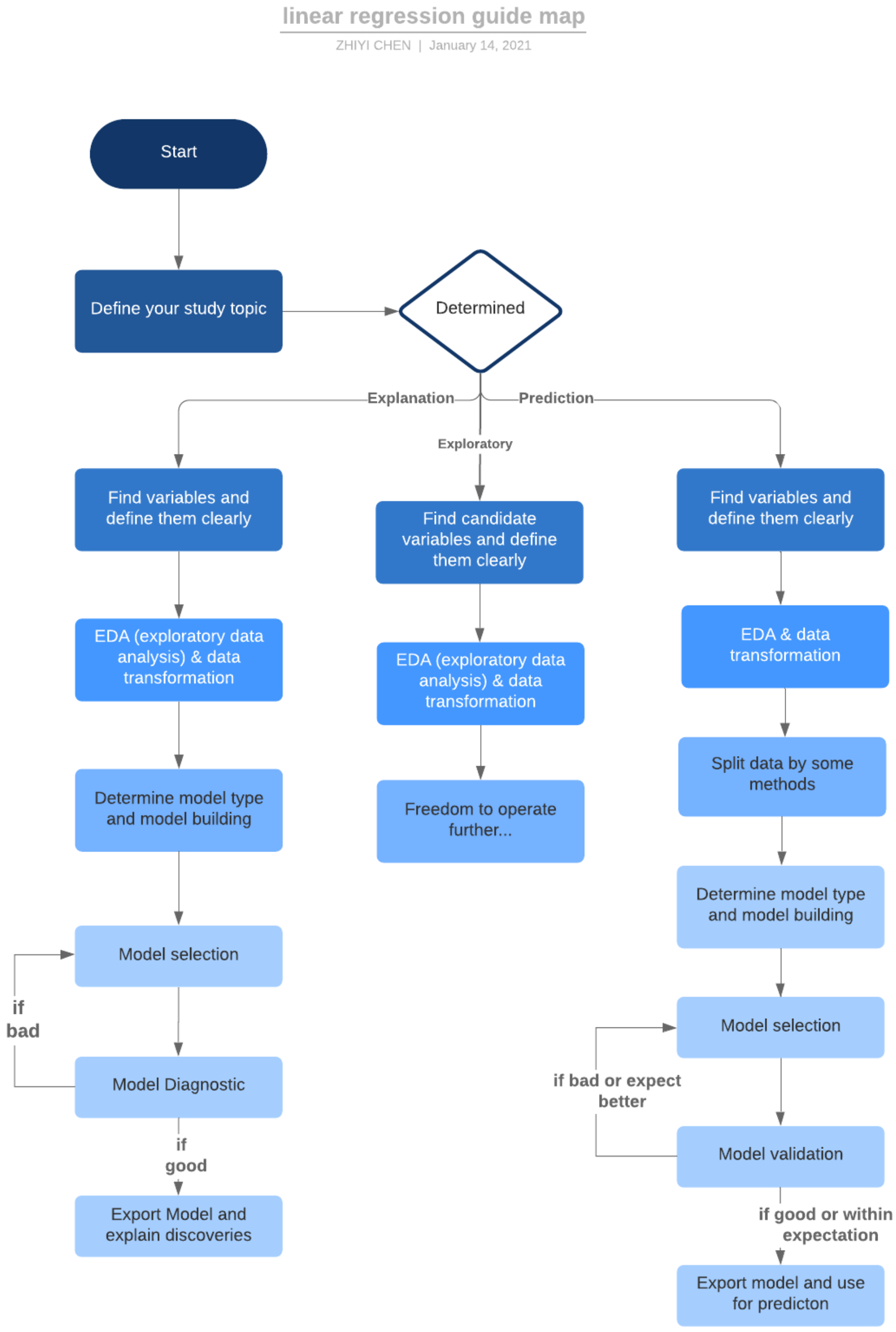

linear regression guide map
ZHIYI CHEN | January 14, 2021
Start
Define your study topic
Determined
Explanation
Prediction
Exploratory
Find variables and define them clearly
EDA (exploratory data analysis) & data transformation
Determine model type and model building
Model selection
if bad
Model Diagnostic
if good
Export Model and explain discoveries
Find candidate variables and define them clearly
EDA (exploratory data analysis) & data transformation
Freedom to operate further...
Find variables and define them clearly
EDA & data transformation
Split data by some methods
Determine model type and model building
Model selection
if bad or expect better
Model validation
if good or within expectation
Export model and use for predicton

Figure 5: Data analysis guide map

2.2 Analysis Topic

All analysis is around our selected topic, I divide topic into three branches

- Explanation: we got response and candidate predictors and we want to see the influence of predictors on response. In real industry problem, we can use explanation method to find significant factors and pay more attentions on these factors.

- Prediction: we got response and candidate predictors, but this time we want to build a model to predict the data of response. In this process, the relationship between predictors doesn't matter much, what we need most is the performance of model validation. In real business, we can predict the future performance of a indicator and build corresponding plan to deal with the potential problems.

- Exploration: we don't know our response or target, but we got a bunch of data and there is enough freedom to select our interest topic to explore. In industry, normally it happens on market analysis where we can select several dimensions to analysis and conclude competition.

Please remember that once we determine our topic, all further operations we might make should have contribution to our topic. We can have our own analysis idea during working process as long as our goal is certain.

2.3 Data & EDA

After determining our path, we will have relevant data. To be honest, how to obtain data is a heavy task since data are scattered which is not our core here. When data are in our hands, firstly we should write a instruction to clearly state the definition and metrics of variables. One more thing should also be emphasized is the data type, qualitative including nominal, ordinal and quantitative including discrete and continuous (we can regard discrete and continuous data as numerical data). In reality, nominal, ordinal and numerical data are main data type we should focus on. Therefore, in EDA(exploratory data analysis), it is necessary to find methods comparing relationships between these data types.

EDA provides a visualized method to discover the pattern behind data which is qualitative, actually we need quantitative way to obtain strong evidence to decide whether to convert data. Before introducing the correlation coefficient, first we should distinguish correlation and dependency [2].

- **Dependency**: A variable whose value depends on the value assigned to another variable (independent variable).

- **Correlation**: The relationship between two or more variables is considered as correlation. The correlation coefficient always assumes *linear relationship* regardless of whether that assumption is correct or not.

To be clear, dependent does not mean high correlation coefficient and high correlation coefficient does not necessarily mean dependent. However, independent must lead to no correlation. no correlation does not lead to independent since correlation only refers to linear relationship. For example, within $Y = X^3 + X^2$ we can see X and Y are uncorrelated but they are dependent. In reality, we will directly dive to the calculation of correlation coefficient to explore the relationship. Normally we know Pearson correlation coefficient, but there are more correlation coefficient methods worth attention because different data type comparison requires us to apply most proper method to obtain the relation.

2.3.1 Pearson correlation coefficient

The mathematical formula was deirved by Auguste Bravais in 1844, introduced by Francis Galton in 1880 and developed by Karl Pearson[9]. The formula in population can be expressed as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

For sample data, it has the same structure with different symbol

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

The assumption or meaning of it should be emphasized

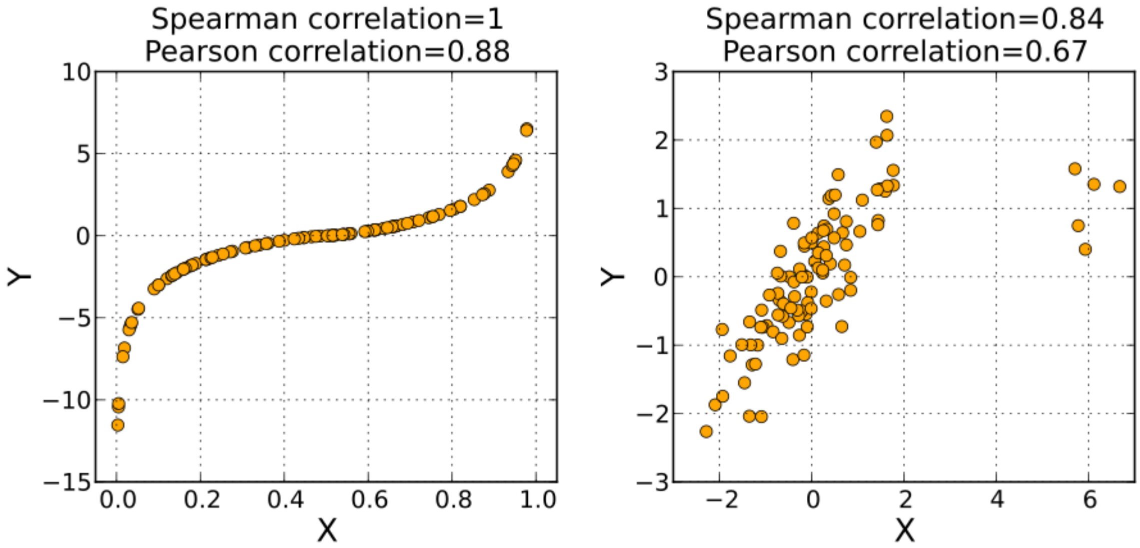
- X and Y can be any type of variables, but if they are nominal or ordinal variables, we should transform them to numeric one. For example, we can convert them to rank value.
- It only used to explain linear relationship.

2.3.2 Spearman's rank correlation coefficient

To calculate this coefficient, we should convert variables to ranking variables. The formula of Spearman's rank correlation is the same as Pearson correlation coefficient's whereas the meaning is different[11].

$$r_s = \rho_{rgX,rgY} = \frac{cov(rgX, rgY)}{\sigma_{rgX}\sigma_{rgY}}$$

where rgX, rgY refer to the ranking variables of X, Y respectively. Spearman's assesses how well the relationship between two variables can be described using a monotonic function rather than explores linear relationship. Few graphs in wiki[11] can explain the difference.



(a) Comparison 1

(b) Comparison 2

Figure 6: Spearman and Pearson correlation comparison

2.3.3 Kendall rank correlation coefficient

Maurice Kendall developed Kendall rank correlation coefficient in 1938 and this coefficient is a statistic used to measure the **ordinal** association between two measured quantities. We don't need to convert variables to ranking ones but we should understand two definitions with observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ of the joint random variables X and Y: if either $x_i < x_j$ with $y_i < y_j$ or $x_i > x_j$ with $y_i > y_j$, then $(x_i, y_i), (x_j, y_j)$ are **concordant**, if $x_i = x_j$ or $y_i = y_j$, then they are **tied**, otherwise they are **discordant**. Then, Kendall τ coefficient is defined as[6]

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{\binom{n}{2}}$$

with assumption that the values of X and Y are unique. However, we can rarely use this formula since x_i and y_i cannot be unique in real data. Thus, there are three different Kendall rank correlation coefficients which can be chosen based on our request. The complete formula can be found on wiki[6] and here we only emphasize the difference between them.

- **Tau-a** is defined without considering tied condition but we allow the tied data appears.
- **Tau-b** takes the tied group into consideration but we assume the scales of ordinal variables are the same.
- **Tau-c** is the most flexible one allowing tied data and different scales.

2.3.4 Variance inflation factor(VIF)

Variance inflation factor is an indicator to justify the multicollinearity level of regressor. It regards the target regressor as response and other regressors as variables, through which we can judge only numerical variables can be used by VIF and we want to know how much variance of target regressor can be explained by others The variance inflation factor of regressor X_k is

$$VIF_k = \frac{1}{1 - R_k^2} \quad k = 1, 2, ..., p - 1$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the p-2 other X variables in the model. Therefore, with our understanding, higher the VIF is, higher colinearity of this regressor has. We can have two rules to deal with the VIF value:

- For VIF of single regressor, "VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable[5]."
- For VIF condition of all regressor, we can treat that $maxVIF_k >> 10$ indicates multicollinearity and $\overline{VIF_k} >> 1$ indicates multicollinearity[8].

To conclude, although Pearson, Spearman's rank and Kendall rank correlation coefficients are defined differently, somehow they can distinguish the relationship between certain variables. Therefore, when we obtain data of **nominal**, **ordinal** and **numerical** types, what is the best or most appropriate way to calculate their correlation, I create a diagram to show my idea.

	Nominal	**Ordinal**	**Numerical**
Nominal	Spearman	Chi-square test/mosaic plot	Scatter plot/-Pearson/VIF
Ordinal		Spearman/Kendall $\tau - c$	Pearson/ Spearman/VIF
Numerical			Pearson/VIF

In my opinion, the most hard part is to compare nominal with ordinal or numerical variables since nominal variable cannot be scaled. Therefore, we can use visualization method combined with quantified formula to deal with it. Please in mind that statistics is full of freedom which allows people to keep as much imagination as possible to solve problems.

After obtaining the coefficient value between variables, next step is a qualitative part. What is the threshold of coefficient value beyond which we should convert data? You can think this question very easy, but also hard as well. If we have former experience or we are confident enough to the threshold we create, it is not a problem. Therefore, asking for a more scientific and statistical question, we might need to use some test, such as a/b test to give a decision.

The last thing to pinpoint is that whether our decision is highly linked to our goal. High correlation coefficient influences the precision of parameter, thus if our goal is prediction, we don't care about the estimation. If we are seeking for explanation of certain variable, then it will be our significant task.

2.4 Model building

Model building is a process of initiation of model selection. Based on the data and data visualization or analysis result, we choose and create first model structure, such as multiple linear functions, exponential functions, logarithmic functions or other models.

The most lucky situation is that through data visualization, it is clear for us to choose a model type which we are familiar with. But when the number of variables is too large or the data visualization cannot give some useful information, what we can do is to give a assumed model. To be honest, multiple linear regression is the most common one. Don't worry, no one knows which is the best model, what we should pursue is the improvement on the model type we choose once we make decision. After we give the optimization, building model of another type can be raised.

Assume we decide to use multiple linear regression, first we need to state the model structure super clearly, especially the assumptions since we need to verify these in model diagnostic part.

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \epsilon,\ \epsilon \sim^{iid} N(0, \sigma^2)$$

with assumptions

- The expectation of Y is linear.
- The errors ϵ are normally distributed.
- The errors ϵ are constant variance (Homoscedasticity).
- The errors ϵ are independent and identically distributed.

Keep in mind that it is just an example above, I can completely define the errors are heterogeneity and they are dependent, but we must make sure there are approaches to verify these properties.

2.5 Model selection

Model selection is a very big and complicated topic and we have lots of methods to create candidate models including value choice of hyperparameters, subset of variable(forward selection, backward selection, forward stepwise), etc. Whatever method we apply, we need a criterion to evaluate each candidate model. In detail, we want to measure the distance between candidate model and true model. Normally a good criterion is able to take the benefit of complexity as well as penalty of complexity into consideration[12]. There are some criterion widely used worth being applied in model selection.

- **Adjusted determination coefficient:** $\overline{R}^2 = 1 - (\frac{n-1}{n-p})\frac{SSE_p}{SST_p}$
- **Mallow's C_p:** $C_p = \frac{SSE_p}{s^2} - n + 2p$
- **Akaiki information criteria(AIC):** $AIC = nlog(SSE_p) - nlog(n) + 2p$
- **Bayesian information criteria(BIC):** $BIC = nlog(SSE_p) - nlog(n) + plog(n)$

where n is the number of observation, p is the number of regressor, SSE_p is the sum of square error under the model with p regressor, SST_p is the sum of square of total error, s^2 is the residual mean square after regression on the complete set of full regressors. The first two criteria are driven from ordinary least square estimation and the latter two are figured out by maximum likelihood estimation. **OLS** and **MLE** can be converted to each other under data with Gaussian distribution, the ideas behind them all make sense. In real problem solving situation, we can choose one of these criteria in model selection. Although they have difference idea and structure, there is little impact on model comparison. As for the concept and statistical difference, I will write another article to address it mathematically.

2.6 Model diagnostic and Model validation

According to our goal (explanation, exploration, prediction), we apply different methods to assess our model. For prediction, we only care the accuracy of model used on test data. Therefore, keeping the mean square error from train data not low and the distance between MSE from test data and train data close is our goal. If not, we need to go back to model building part and reconstruct our model.

For explanation and exploration, we use model diagnostic, that is, to make sure our model conforms to the assumptions and has high fitness to data. Therefore, I divide diagnostic into two parts, assumption verification and outlier justification.

Assumption verification mainly faces the linearity, normality and Homoscedasticity and we can use visualization plots to show them, such as scatter plot for linearity, QQ plot for normality, residual plots for linearity, normality and Homoscedasticity. Here we need attention that in order to obtain more clear and precise residual plot, we can have transformation for residuals, such as **studentized residuals**, **studentized deleted residuals**[13], etc. A visualization example showing the model is not linear.

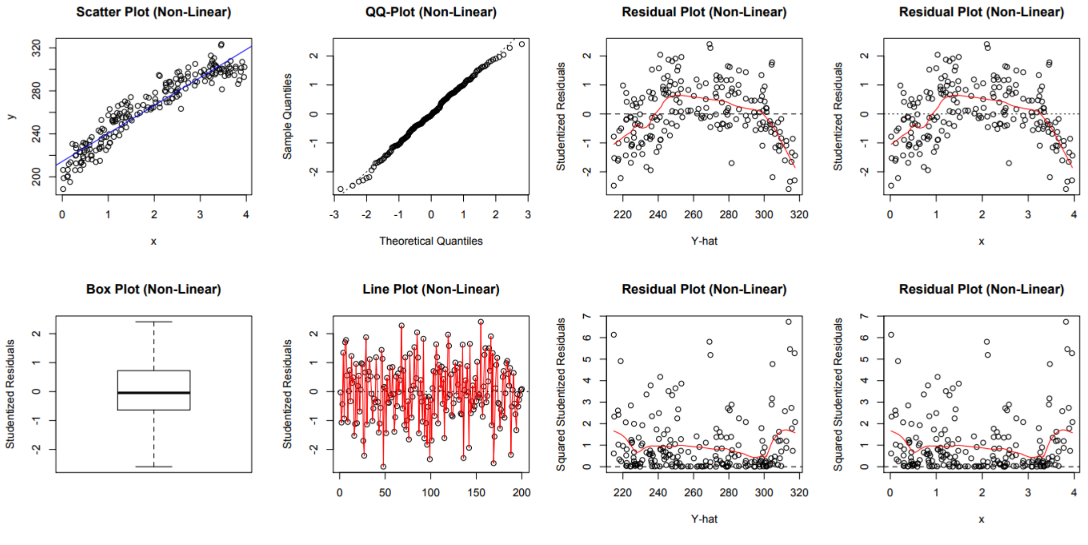


Figure 7: Diagnostic Plots

From the residual plots, non-linearity is obvious which leads us back to model building and selection part and redefine our model getting more fitness to observations. Here we finish a cycle and iterate models until we are satisfied.

Outlier justification aims to fix the influence brought by outliers. During model estimation, outliers has no right involved but it is hard to be avoided. In this article[8], we have many statistical methods to distinguish outliers and next step is the same as previous one that after thinking about whether these outliers matter, we should make decision whether choose more proper model.

3 Conclusion

This article mainly focuses on the basic concepts understanding before model building, such as random variable, degree of freedom, confidence interval, etc and the procedure when solving problem using model. However, there are large number of details worth being researched.

Sampling is an inevitable problem in real life. How to sample dealing with different data, is there any method to evaluate the quality of sample, how to sample in several tests, such as A/B test? It is a necessary road and foundation to statistics and model.

Secondly, the model selection criterion comparison is very interesting which displays the subtle idea of statisticians. How did they come up the idea, is it highly related to the data we choose, is there a rule to guide us when to choose which criterion? All of these are full of value.

Anyway, statistics is a "free" subject, the guideline mentioned in this article is just my idea, everyone can have their map to solve problem. All I always believe is studying from people (learning from statisticians) and exploring from data.

References

[1] Jason Brownlee. “What is the Difference Between a Parameter and a Hyperparameter?” In: (2017). URL: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/#:~:text=In.

[2] *Correlation and Dependency.Aren’t they One and the same?* URL: https://www.jigsawacademy.com/correlation-and-dependency-arent-they-one-and-the-same/#:~:text=Dependency.

[3] *Degree of freedom.* URL: https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)#:~:text=In.

[4] Minitab Blog Editor. “What Are Degrees of Freedom in Statistics?” In: (2016). URL: https://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-degrees-of-freedom-in-statistics.

[5] Jim Frost. *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions.* URL: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/.

[6] *Kendall rank correlation coefficient.* URL: https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient.

[7] Will Kenton. “Random variable”. In: (2020). URL: https://www.investopedia.com/terms/r/random-variable.asp#:~:text=Key.

[8] *Measures of Influence.* URL: https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html.

[9] *Pearlson correlation coefficient.* URL: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

[10] *Random variable.* URL: https://en.wikipedia.org/wiki/Random_variable.

[11] *Spearman's rank correlation coefficient.* URL: https://en.wikipedia.org/wiki/Spearman%5C%27s_rank_correlation_coefficient.

[12] Marco Taboga. *Model selection criteria.* URL: https://www.statlect.com/fundamentals-of-statistics/model-selection-criteria.

[13] Gabrial young. "Diagnostics and Remedial Measures". In: *GR5205 Linear Regression Model* (2019).