# Week 5: Linear Regression

* basics of regression (we’ll cover advanced regression topics later in the course),
* along with the concepts of estimating a model’s quality and distinguishing between correlation and causation.

# Additonal notes

https://www.youtube.com/watch?v=vGOpEpjz2Ks

box-cox

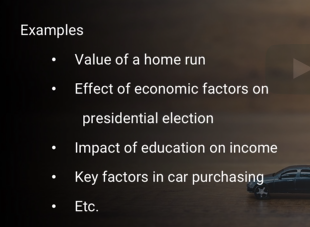
library faraway (butterfat - dataset)

mass(for lm)

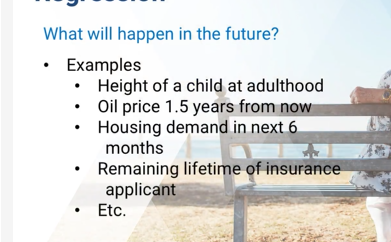
# 5.1 Simple Linear Regression (SLR)

## Answers below questions (Most used models)

1. how do system work? (descriptive)



2) Prediction : what will happen? (predictive)



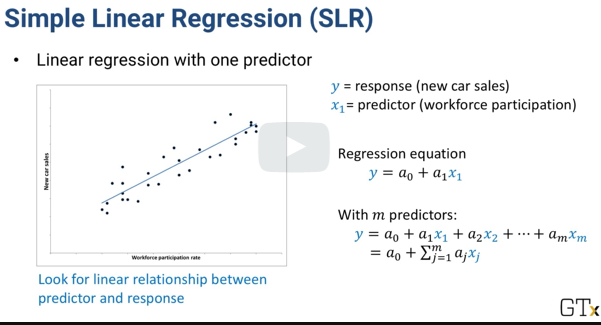
## Simple Linear regression(SLR)

Simplest is Linear regression with one predictor (how the predictors influence the decision)

-more people working..more car sales…so they are related..(linearly related)

Regression : identifies how much influence?

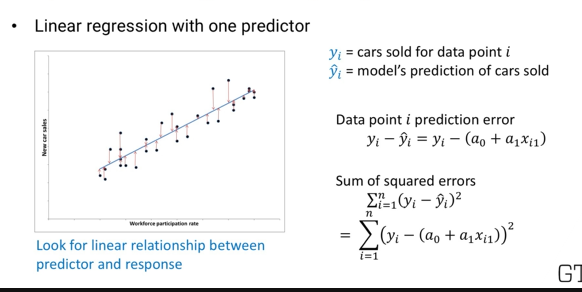
Regression is a line that that goes through set of points



How to measure quality of the fit? We can use sum of squared errors

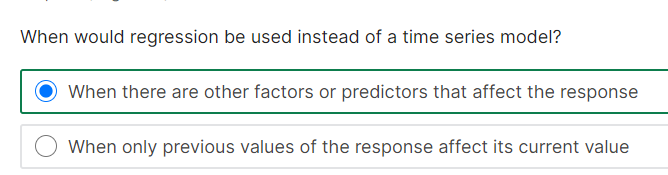
SSE : distance between true number of cars sold vs. linear regression line

-in below equation, we replace the line witth the above equation



**Best Fit regression line:** values of co-efficients a(0) and a(1) that minimizes the sum of square error

Math : SSE is a convex function that we are trying to minimize. So we take partial derivative of SSE with respect to each constant and set them to Zero. Solve the equation simultaneously to find the minimum SSE



# 5.2 Maximum Likelihood and Information Criteria

**Agenda**: measure the quality of a model’ fit

## Basic measurement: Likelihood

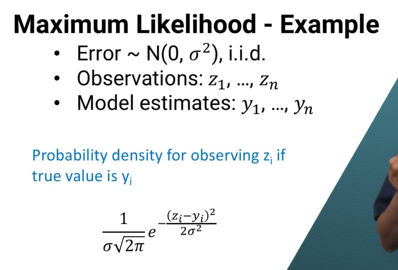
**Assumptio:** observed value is correct value ; we already have info about the variance

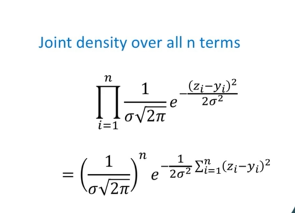
**Steps:**

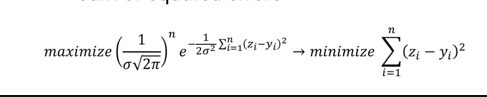
- For set of parameters , we measure the probability density function

- whichever model gives the maximum probability density function (PDF), is the maximum likelihood

**Example:**

* Errors are normally distributed
* Mean =0;variance=sigma square (independent from one data point)
* Below is the probability of observing z(i) for the y(i)
* 
* Since the errors are independent, joint probability is below for 1 through “n” z values(observations\_

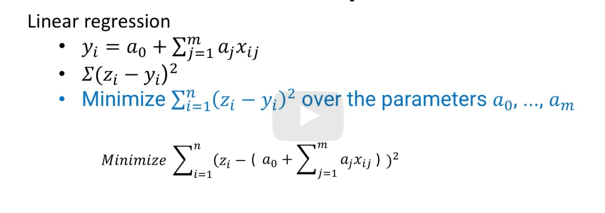


* As the expontial term gets larger as exponent gets larger, we can find largest value of this expression (by finding the largest exponent)
* In short, if errors are normally distributed and independent, we can remove constants , which will give us sum of squared errors. So whichever parameter minimizes SSE is maxiumum likelihood
* 

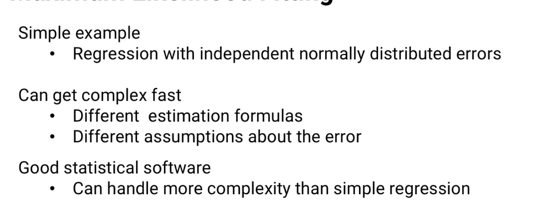
**Objective for linear regression**

Z= > observation

Y => model estimates(which was straight line) that we can substitute straight line equation



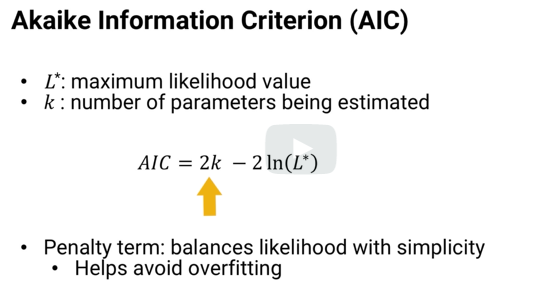
* Using Optimization above is a easy problem to solve.
* We can use likliehood to compare 2 models by conducting hypothesis testing.
* Regression with independent normally distributed error is a example of maximum likelihood fitting

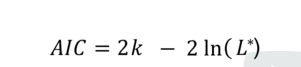
Ak

## Akaike Information criteria (AIC)

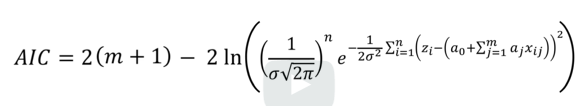
Function that combines maximum likelihood with model complexity

* This function is applied so we can keep the number of parameters used to attain maximum likelihood in check to avoid overfitting (to retain the simplicity)
* K => penalty term



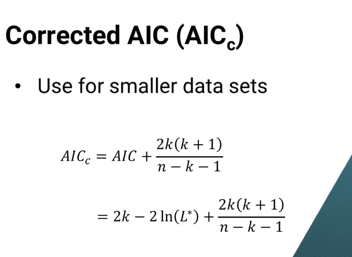


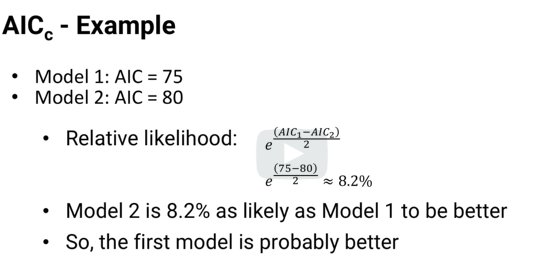
In above maximum likelihood, replace L with maximum likelihood formula and “k” with number of parameters(m+1)



-Obejctive: model with **smaller AIC is preferred**

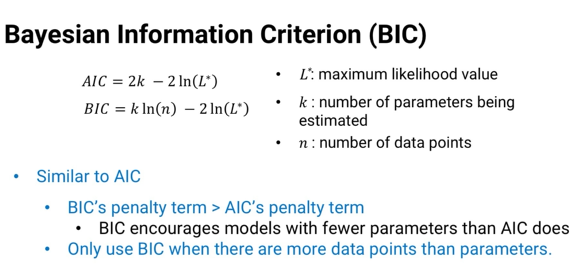
**Note AIC is good for inifinite data point.** For practical purpose, we use “correction term”



* When comparing 2 models, we can get relative likelihood (lower AIC model is better)
* 

## Bayesion Information criteria (BIC)

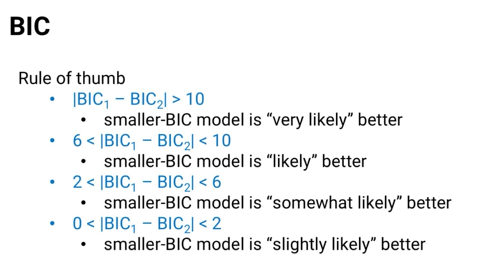
-similar to AIC



**Notes**

* penalty is larger than AIC
* BIC encourage models with fewer parameters (when number of parameters “k” gets closer to number of data points, it breaks)
* INEXACT SCIENCE

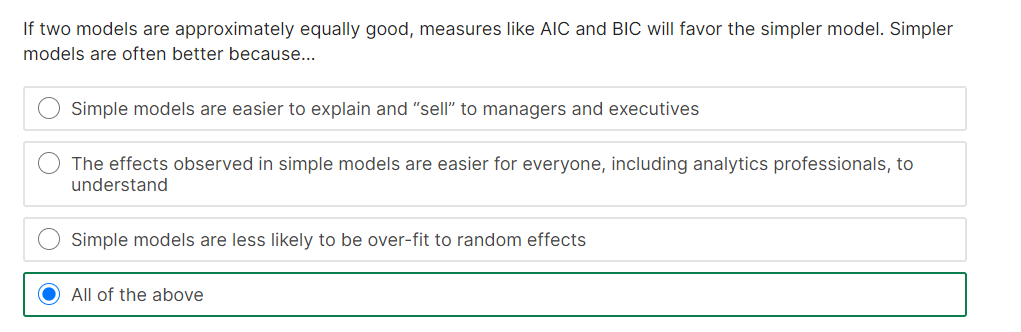
**COMPARING DATAMODELS: When comparing two data models using BIC, here is the rule**

****

**NOTE**

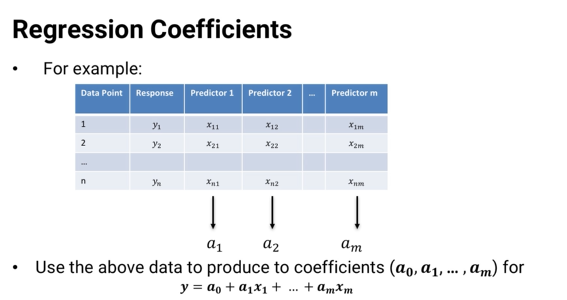
**AIC is frequentist point of view**

**BIC is Bayesian point of view**

****

# 5.3 Using Regression

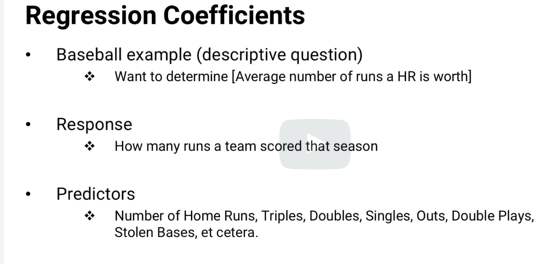
**Agenda**: interpret regression co-efficients



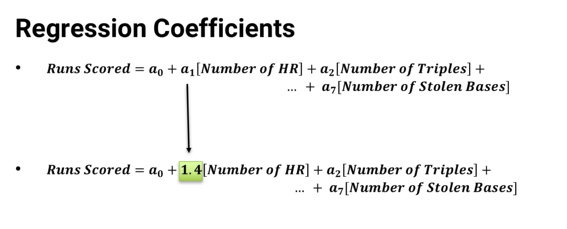
## Descriptive(what is happening)

**Key answers**: co-efficient

**Sample**:



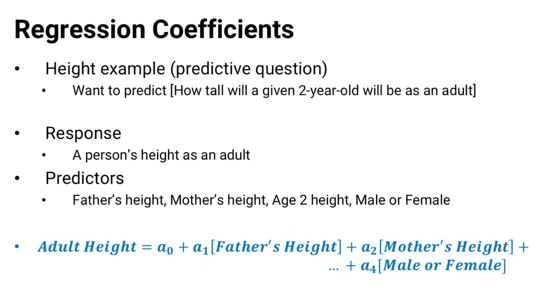
HR => Home runs



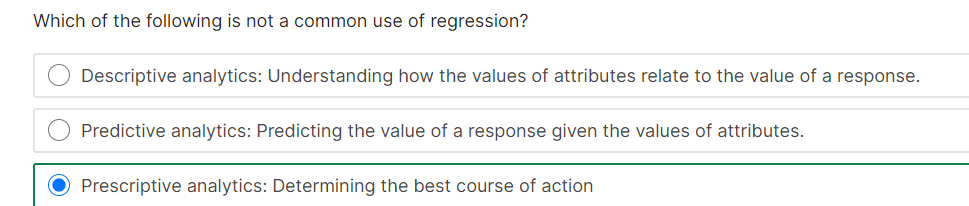
Meaning every home run will add “1.4” runs to the team’ total

## predictive(what is happening)

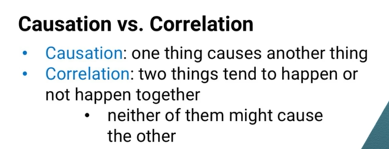
**key answers: predicted responses**



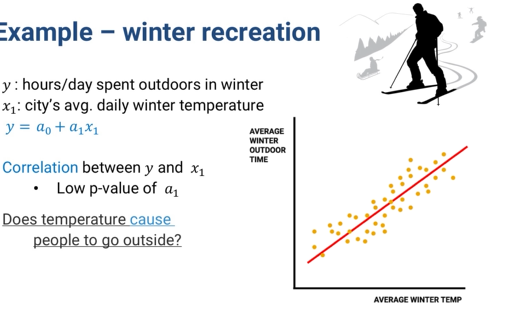
\* if given a new data point, say, 2 year old boy and height of father and mother, use this linear equation to predict the “response”



# 5.4 Causation vs Correlation

****

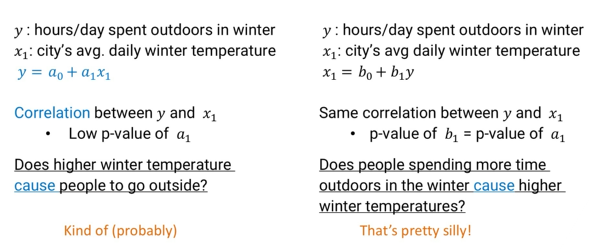
## Causation

****

**Causation :** Will high weather in winter result in more people outdoor? Maybe

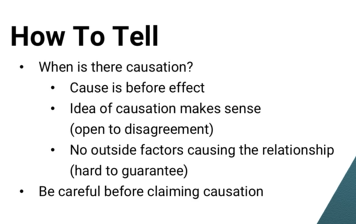
## Co-relation

**Correlation:** Vice versa..Does more people outdoor cause high weather in winter ? NOT (The data amy be correlated , but one doesn’t cause other)

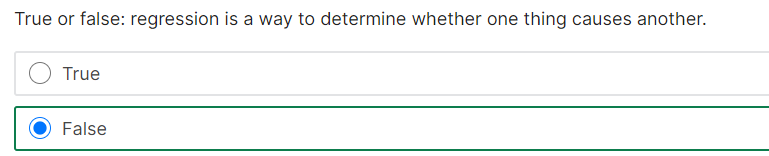
****

But we can still use correlation to predict weather using historical data of number of outdoor days in winter to predict weather in future

## How to find “CAUSE”? ( 8 rules of thumb)



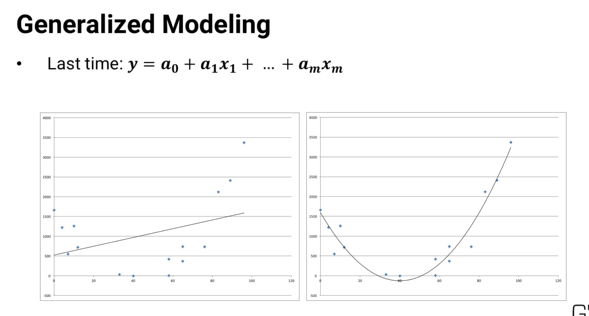
****



# 5.5 Transformation and Interactions

**Agenda:**

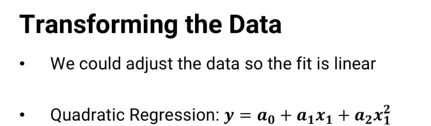
1. Build general regression
2. Interpret output
3. What is important and what is not



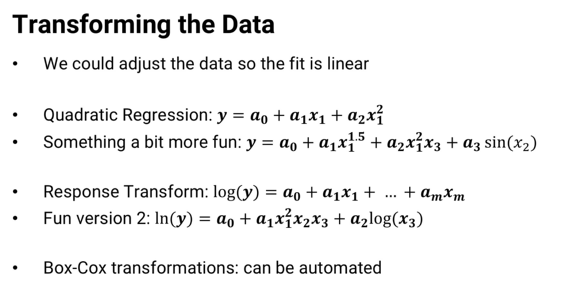
## Transform data:

If data looks quadratic, we can include a quadratic equation to fit the data(Create new column of data)

* We can transform the response or transform both data and response



More samples

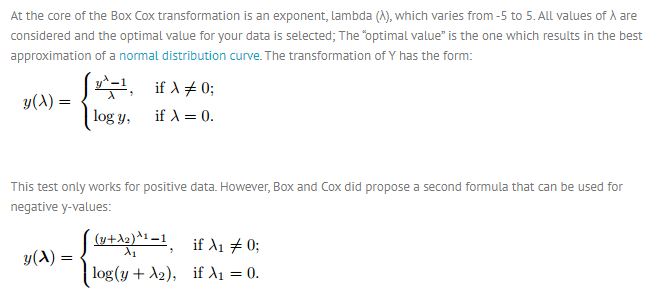


Box Cox transformation can be automated by statistical software

**What is Box Cox :** two scientist with name Box and Cox.

A Box Cox transformation is **a transformation of non-normal dependent variables into a normal shape**.

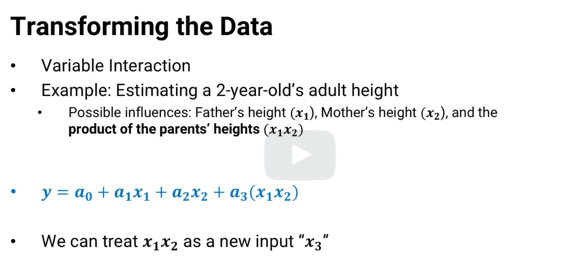
They use a lamba function

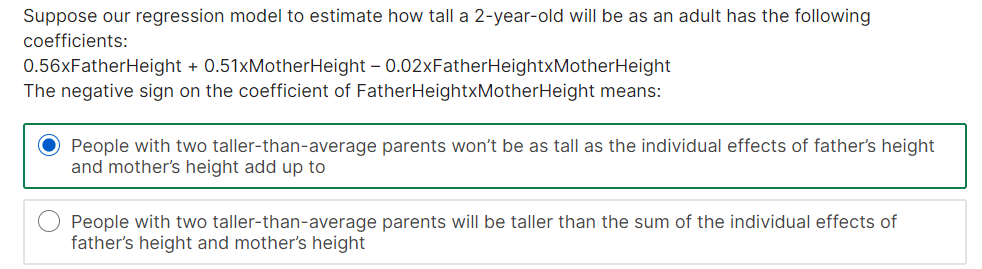


Interaction Term

-use Interact term to the existing data to better show the relation. In samle below, the “Y” is dependant on product of X(1) and x(2).

- Once the interacted term is added as a new column of data, it can be transformed and used in the regression as any other predictor





# 5.6 Regression Output

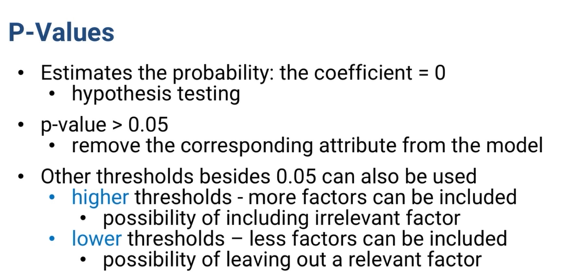
**Agenda:**

* Look at output to determine the significance of the co-efficient

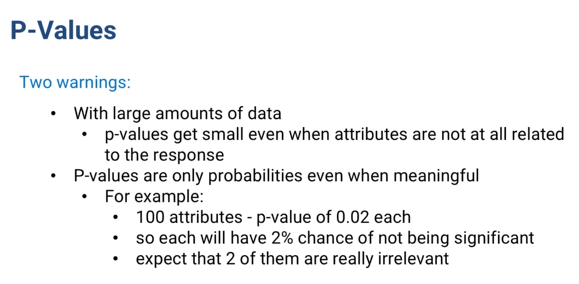
## How to determine?

## P-value

* p-value for co-efficients (estimates that the c0o-efficients really might be zero;hypotheses testing)



### Warning of using p-value



## 2. Confidence Interval (CI)

- most software will give 95% CI

- determine importance of co-efficients related to p-value (How close it is to Zero)

## 3. T-statistics

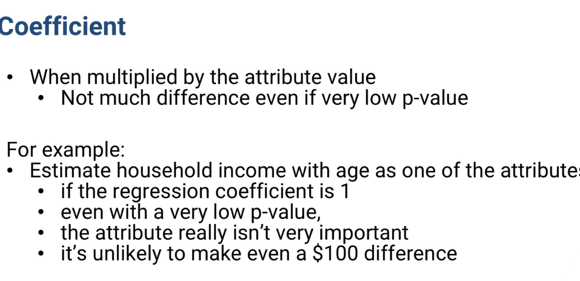
- co-efficient/ standard error

* related to p -value

## 4. Co-efficients itself

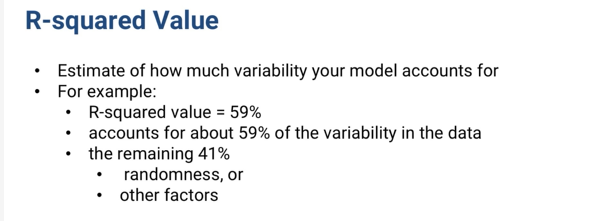
Co-efficient \* attribute value doesn’t make difference (Even with low p-value).

If regression co-efficient is 1 even with low p-value, the attribute if not important



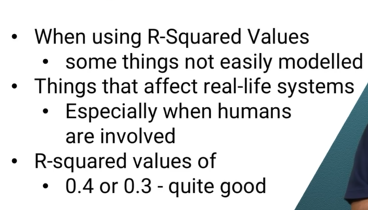
## 5. R-square value (co-efficient of determination)

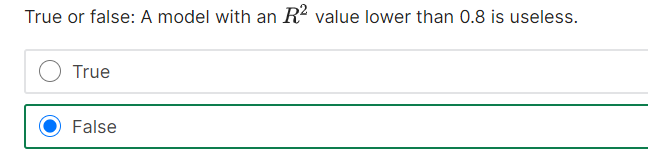
This is a model metrics that accounts for variability of model



We also have “Adjusted R Square” that accounts for **number of attributes used**

**Summary: (Objective: more R squared, more better)**



****