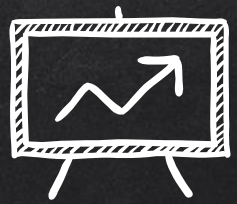


# Regression and Prediction

– Lucy Xu

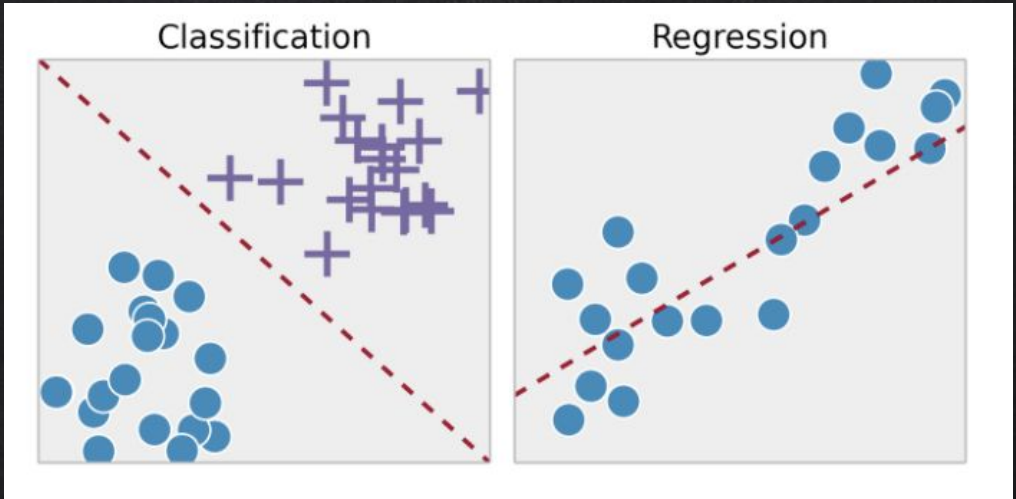
# Quiz

- ✘ What is Simple Linear Regression(MLR) assumptions?
- ✘ What is Cross Validation and why/how to do it?
- ✘ What's the logistic regression loss function?
- ✘ What's gradient descent?
- ✘ What's regularization?

# Types of Machine Learning

- ✘ Supervised Learning
  - Target type: Continuous and Classification
  - Algorithms: Linear regression, logistic regression, Decision Tree, Random Forest, XGB, Naive Bayesian, SVM, Neural Network, KNN
- ✘ Unsupervised Learning
  - Algorithms: K-means, PCA

# Typical Problems of Supervised Learning




Decision: Approve/Decline,
Solicitation/not, market/not,
promotion/not;
Attrition: Leave/Stay,
happy/unhappy
Risk: Yes/No

House Value,
Product price,
Customer income

# Linear Models

A model which is linear in the parameter

SLR model: $y = \beta_0 + \beta_1 x + \varepsilon$

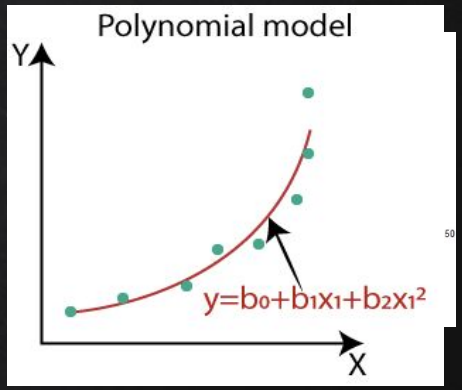
MLR model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Polynomial regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_{1.squre} + \varepsilon$$

Nonlinear model example:

Logistic growth model: $y = \alpha/\beta e(-kx) + \varepsilon$



# Topics to covered in Regression

- ✘ Simple linear regression
- ✘ Linear algebra 101
- ✘ Multiple-regression
- ✘ Logistic regression
- ✘ Gradient descent
- ✘ Regularization


# Simple Linear Regression

"This algorithm assumes that there is a linear relationship between the 2 variables, Input (X) and Output (Y), of the data it has learnt from. The Input variable is called the *Independent Variable* and the Output variable is called the *Dependent Variable*. When unseen data is passed to the algorithm, it uses the function, calculates and maps the input to a continuous value for the output."

For example: If X increase Y will increase to a certain magnitude or decrease to a certain magnitude.

# Linear Regression Key Terms

**Dependent variable** – The variable is to predict. Also called: Response, Y variable, target, outcome

**Independent variable** – The variable used to predict the response. Also called: X variable, feature, attribute, predictor

**Observation** – The vector of predictor and outcome values for a specific individual or case. Also called, row, case, instance, example, record, $(x_1, y_1)$, $(x_2, y_2)$....

**Intercept** – The intercept of a regression line is the predicted value when X=0 . Also called $b_0$, $\beta_0$,
**Regression coefficient** – The slope of a regression line. Also called, slope, $b_1$, $\beta_1$, w1

**Fitted values** – The predicts obtained from the regression line. Also called, estimates, predicted values, $\hat{Y}_i$

**Residuals** – The difference between the observed values and the fitted values. Also called, errors **Least squares** – The method of fitting a regression by minimizing the sum of squared residuals. Also called, ordinary least squares(OLS, MSE)

**Maximum Likelihood Estimation** – "a method of estimating the parameters of a probability distribution by maximizing a likelihood function", also called, MLE

# Simple Linear Regression Equation

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\hat{Y} = b_0 + b_1 X$$

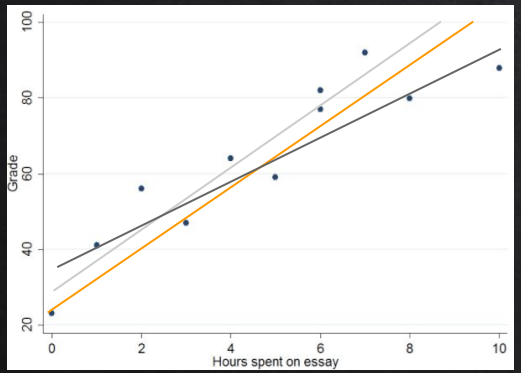
Simple Linear regression assumption:

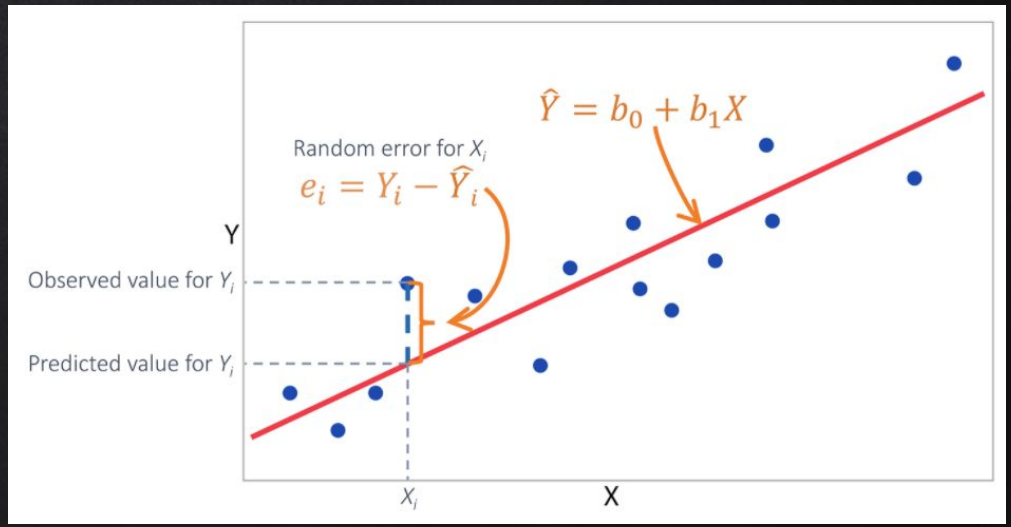
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i=1,2,3,4,.....$

1. Error is zero on average
2. Error has constant variance
3. Each observation is independent to other
4. $\varepsilon_i$ are normal distributed

# Least Squares Estimation

With data $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$...., there may be several ways to fit a linear regression model



# Least Squares Estimation

With data $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$...., find the candidate model($\hat{Y} = b_0 + b_1 X$) which minimize:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2.$$

By mathematical deviation to find the value of $b_0$ and $b_1$ to make S the smallest of all candidates:

Intercept: $b_0 = \bar{y} - b_1 x\_mean$

Slope: $b_1 = r\frac{s_y}{s_x}$

Correlation Coefficient: $r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$

# OLS derivation detail

$$S(\beta_0,\beta_1)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

β0 and β1 are random variables; b0, b1 are the final values to minimize the square of error

The partial derivatives of S(β0,β1) with respect to β0

$$\frac{\partial S(\beta_0,\beta_1)}{\partial\beta_0}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)$$

The partial derivatives of S(β0,β1) with respect to β1

$$\frac{\partial S(\beta_0,\beta_1)}{\partial\beta_1}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)x_i$$

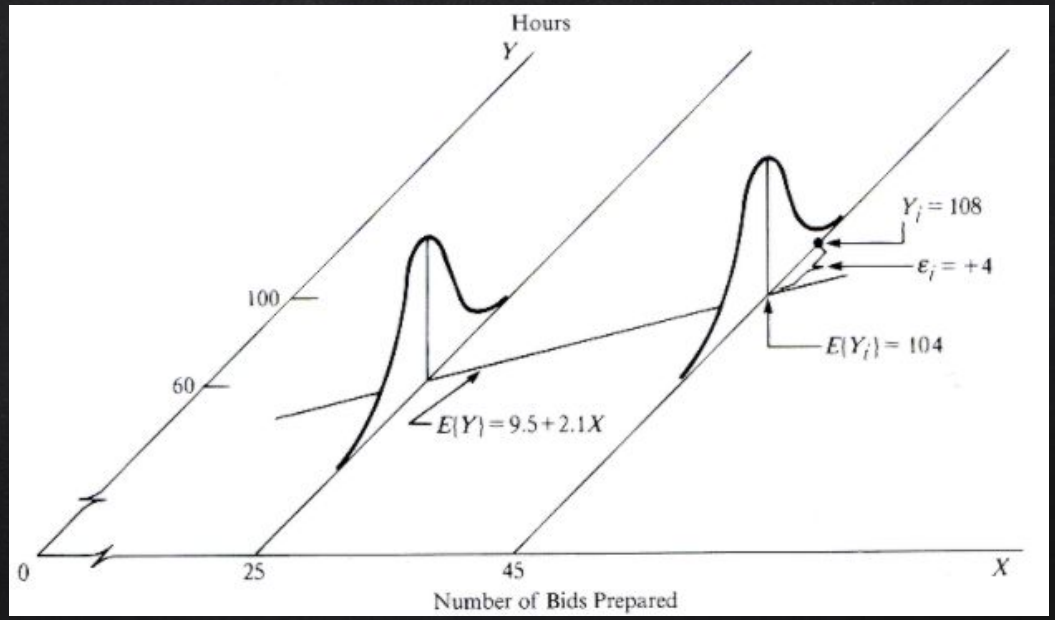
The solution(b0, b1 ) of β0,β1 when S is smallest are obtaining by setting,

$$\frac{\partial S(\beta_0,\beta_1)}{\partial\beta_0}=0$$

$$\frac{\partial S(\beta_0,\beta_1)}{\partial\beta_1}=0.$$

Or through Gradient Descent

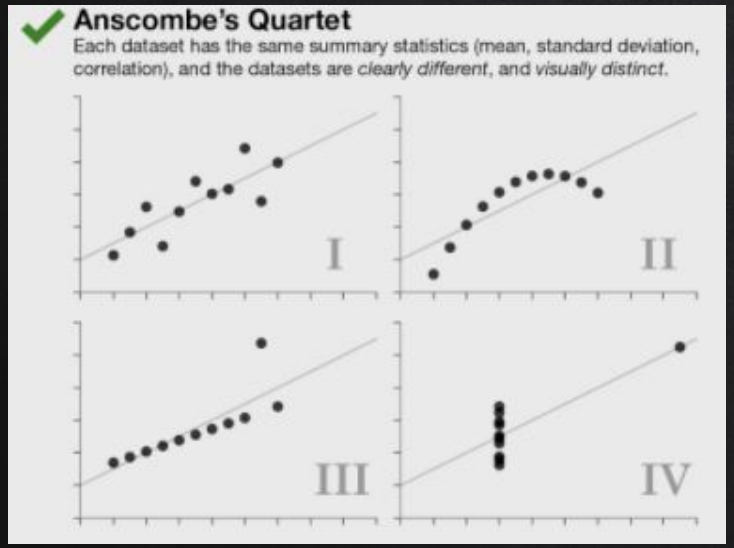
# Revisit the Normal assumption




T test in LR: The coefficient b1 for a predictor, divided by the standard error of the coefficient b1, giving a metric to compare the importance of variables in the model.

# Residual Diagnostic

Regression use for prediction and explanation, your model should be intuitively reasonable beside with the mean error square (MSE)

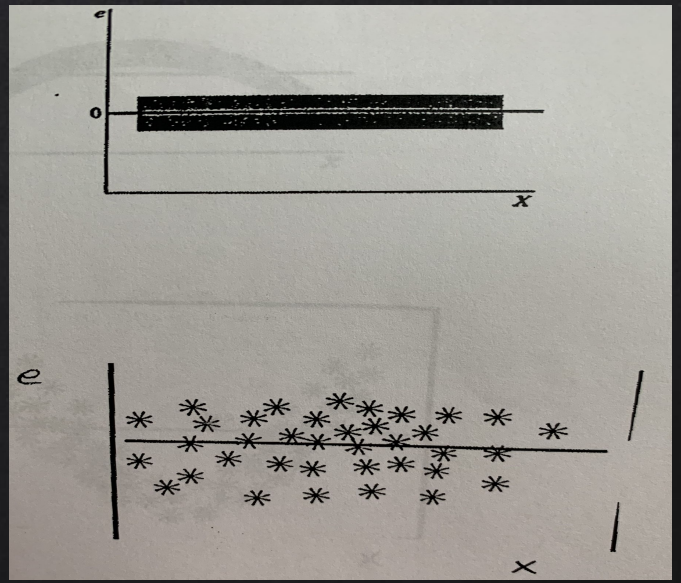



Always conduct a scatter plot for regression

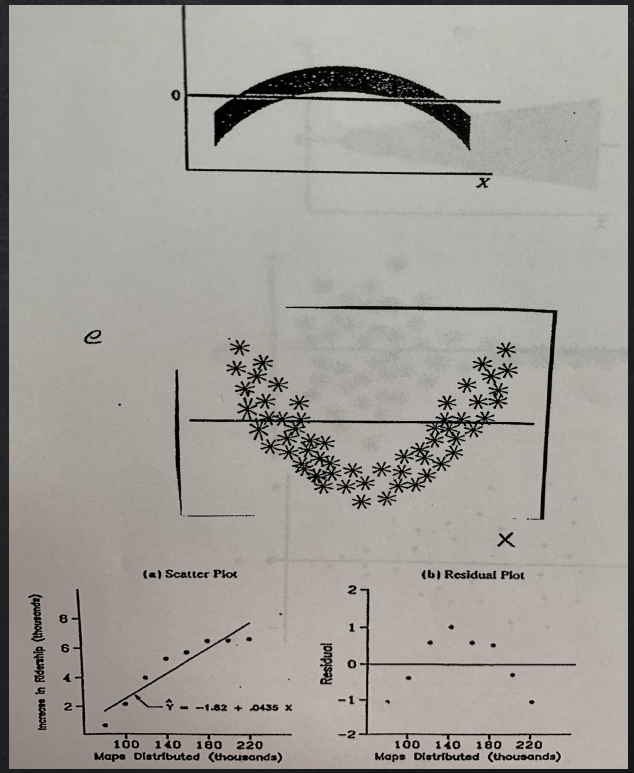
Residual diagnostic help disclose wrong inception and adjust business understand

# Residual Diagnostic – Idea Residuale

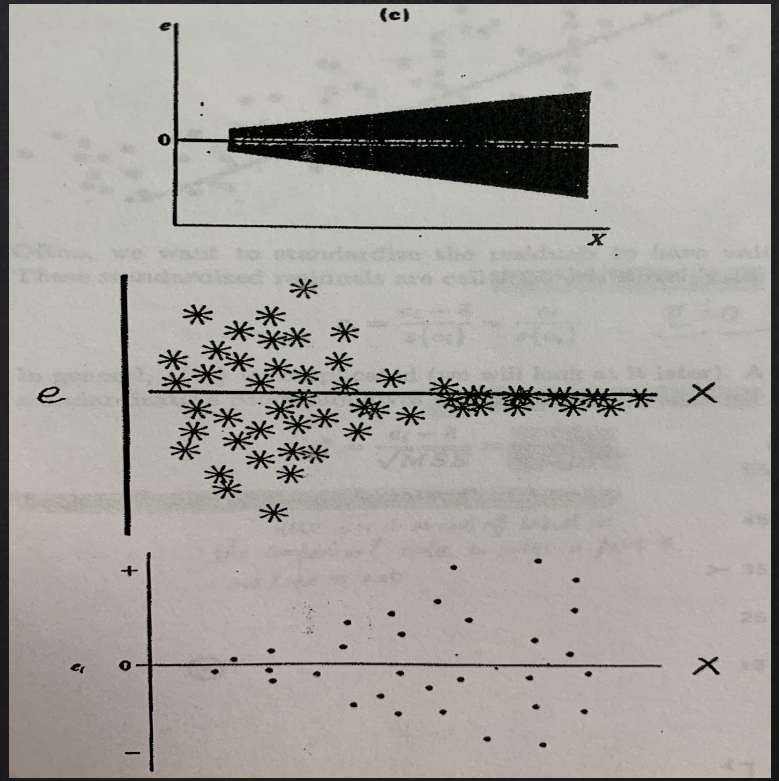
$E_i = Y_i - \hat{Y}_i$, idea residual plot is like this:



# Residual Diagnostic – Trended residuals

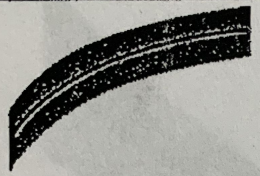
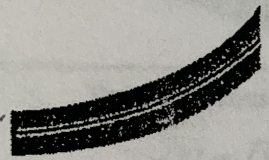
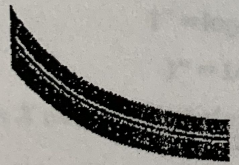


# Residual Diagnostic – Inconsistent Variance

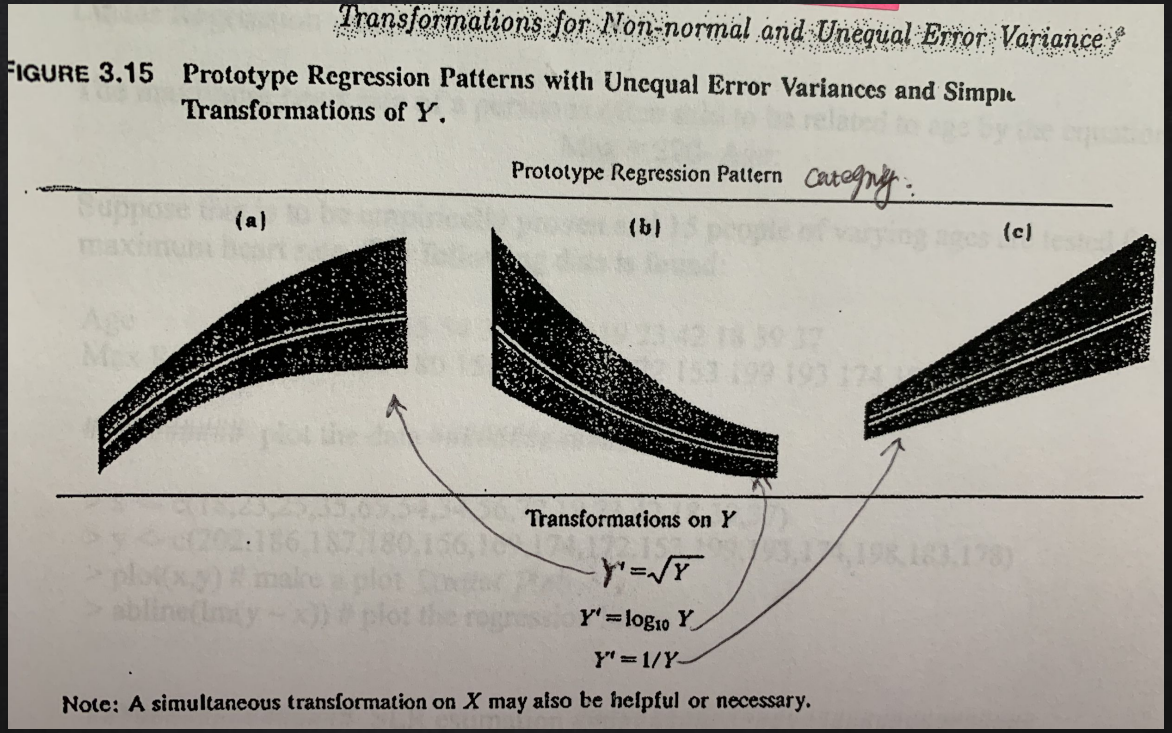


# Independent Variable transformation

**Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of $X$.**

| Prototype Regression Pattern | Transformations of $X$ |
|---|---|
| (a)  | $X' = \log_{10} X$ $X' = \sqrt{X}$ |
| (b)  | $X' = X^2$ $X' = \exp(X)$ |
| (c)  | $X' = 1/X$ $X' = \exp(-X)$ |

# Dependent Variable transformation



*Transformations for Non-normal and Unequal Error Variance*

**FIGURE 3.15** Prototype Regression Patterns with Unequal Error Variances and Simple Transformations of $Y$.

Prototype Regression Pattern

(a) (b) (c)

Transformations on $Y$

$Y' = \sqrt{Y}$

$Y' = \log_{10} Y$

$Y' = 1/Y$

Note: A simultaneous transformation on $X$ may also be helpful or necessary.

# Model Performance Evaluation

MSE: Mean square error

R-square(R2): R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale)

**决定系数**（英语：coefficient of determination，记为$R^2$或$r^2$）在统计学中用于度量因变量的变异中可由自变量解释部分所占的比例，以此来判断统计模型的解释力。[1][2][3]

对于简单线性回归而言，决定系数为样本相关系数的平方。[4]当加入其他回归自变量后，决定系数相应地变为多重相关系数的平方。

假设一数据集包括$y_1,...,y_n$共$n$个观察值，相对应的模型预测值分别为$f_1,...,f_n$。定义残差$e_i = y_i - f_i$，平均观察值为

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

于是可以得到总平方和

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

回归平方和

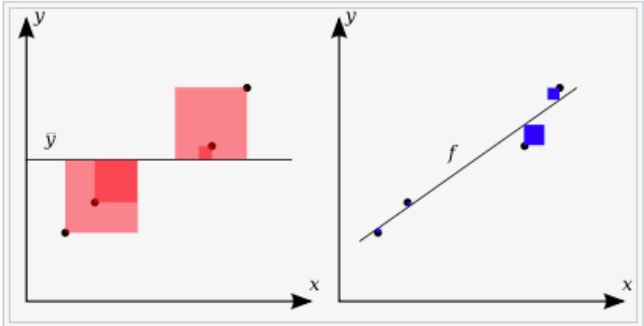
$$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2,$$

残差平方和

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

由此，决定系数可定义为

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$




决定系数$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$示意图 线性回归（右侧）的效果比起平均值（左侧）越好，决定系数的值就越接近于1。 蓝色正方形表示线性回归的残差的平方， 红色正方形数据表示对于平均值的残差的平方。

# Main Takeaway

1. "The regression equation models the relationship between a response variable Y and a predictor variable X as a line."
2. "A regression model yields fitted values and residuals predictions of the response and the errors of the predictions."
3. "Regression models are typically fit by the method of least squares."
4. "Regression is used both for prediction and explanation."
5. A good model(Robust/application relevant) and a reasonable explanation need fully model inspection


# Linear Algebra 101

Quick review of basic Linear Algebra

Two docs here:

Doc1

Doc2


# Multiple-regression

"This is a simple expansion of simple linear regression from 1 predictor to multiple predictors"

$$Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

# Model assumption and Matrix presentation

Multiple linear regression assumption:

1. Error is zero on average
2. Error has constant variance
3. Each observation is independent to other
4. $\varepsilon_i$ are normal distributed

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{2,1} & X_{3,1} \\ 1 & X_{2,2} & X_{3,2} \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & X_{2,n} & X_{3,n} \end{bmatrix} \times \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X} \times \mathbf{B} + \mathbf{e}$$

# Least Square Estimation

$$Q(\underline{b}) = \sum e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots b_{p-1} x_{i,p-1}))^2$$

$$\sum e_i^{\,2} = \mathbf{e'e} = \mathbf{Y'Y} - \mathbf{B'X'Y}$$

Solution: Find the B that minimize Q(b)

$$\mathbf{B} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}$$

$$\mathbf{X'} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ X_{3,1} & X_{3,2} & \cdots & X_{3,n} \end{bmatrix} \qquad \mathbf{Y'} = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix}$$

$$\mathbf{B'} = \begin{bmatrix} B_1 & B_2 & B_3 \end{bmatrix}$$

$$\hat{\sigma}^2 = \frac{\mathbf{e'e}}{n-k} = \text{residual variance}$$

# Cost Function

$$S(\beta_1,\beta_2,\beta_3) = \frac{1}{2N}\sum_{i=1}^{n}(y_i - (\beta_1 + \beta_2 x_2 + \beta_3 x_3))^2$$

## Gradient Descent/partial derivatives

$$S'(\beta_1) = -(y - (\beta_1 + \beta_2 x_2 + \beta_3 x_3))$$

$$S'(\beta_2) = -x_2(y - (\beta_1 + \beta_2 x_2 + \beta_3 x_3))$$

$$S'(\beta_3) = -x_3(y - (\beta_1 + \beta_2 x_2 + \beta_3 x_3))$$

# Accessing Model

Every Machine Learning method/Algorithm should have at least one objective function to optimize upon. And the metrice ties to this objective function will be the key performance indicator(KPI),

For linear model the most important KPI RMSE(Root Mean Square Error), which is the square root of the average squared error. This measures the overall accuracy of the model and is a basis for model comparison/selection

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Other metrics: R 2, RSE, F-statistics, AIC, BIC

# Cross-Validation

**"Data Scientist/statistician beats their data to death to get what they want"**

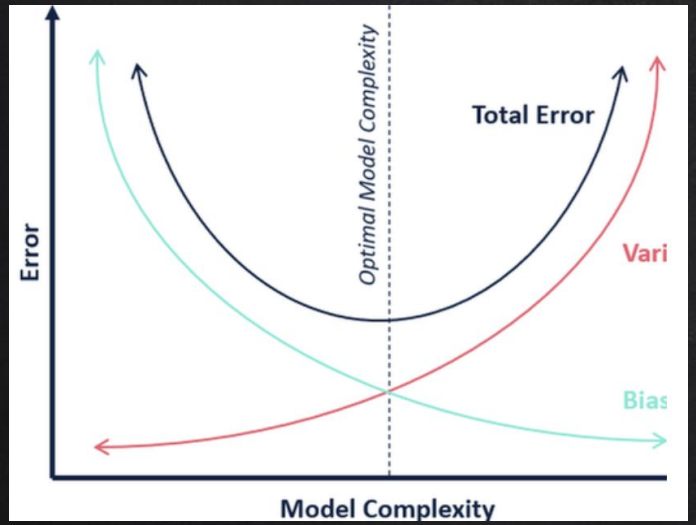
**From certain point of view, it is a true statement. If a data was excessively used to fit a result, it is very easy to encounter another problem – Overfitting**

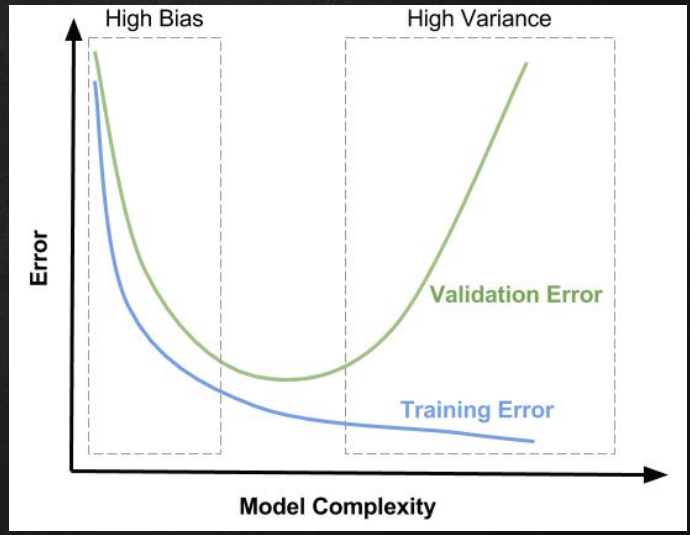
**What's the solution: holdout sample (build, test, validation), n-fold cross validation**

Cross-validation extends the idea of a holdout sample to multiple sequential holdout samples. The algorithm for basic k-fold cross-validation is as follows:

1. Set aside 1/k of the data as a holdout sample.
2. Train the model on the remaining data.
3. Apply (score) the model to the 1/k holdout, and record needed model assessment
4. Restore the first 1/k of the data, and set aside the next 1/k (excluding any records that got picked the first time).
5. Repeat steps 2 and 3.
6. Repeat until each record has been used in the holdout portion.
7. Average or otherwise combine the model assessment metrics.

# Bias-Variance trade-off




1) High Bias – underfitting
2) Goldilocks Zone – just right
3) High Variance – overfitting

# Model Selection

A data may have many independent variables could be used for modeling, you may build the model with a handful predictors, but not necessary means the more variable you add to your model the better model you will get. If the model performance are similar, a simpler model is more preferred.

More predictor the smaller RMSE or high R-square, therefore adjusted R2, AIC, BIC, F-statistics, used to select the best model with minimum predictors

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-P-1}$$

Algorithm used to minimize AIC, BIC, maximize adjusted R2 include backward selection, stepwise regression

# Weighted Regression

Weighted regression used for variety of purposes;
in particular, it is important for analysis of complex surveys.
Data scientists may find weighted regression useful in two cases:

1. Inverse-variance weighting when different observations have been measured with different precision; the higher variance ones receiving lower weights.
2. Analysis of data where rows represent multiple cases; the weight variable encodes how many original observations each row represents.

# Key Takeaway

1. “Multiple linear regression models the relationship between a response variable Y and multiple predictor variables “.
2. “The most important metrics to evaluate a model are root mean squared error (RMSE/MSE) and R-squared ( R 2 ).”
3. “The standard error of the coefficients can be used to measure the reliability of a variable’s contribution to a model.”
4. “Stepwise regression is a way to automatically determine which variables should be included in the model.”
5. “Weighted regression is used to give certain records more or less weight in fitting the equation.”


# Logistic regression

"This is analogous algorithm to MLR with the target is binary "

For example: If X increase the odds of Y=1 compare to y=0 will increase by a certain ratio or decrease to a certain ratio

# Logistic Function

$$p = \beta 0 + \beta 1\, X + \beta 2\, X2 + \ldots + \beta p\, Xp$$

$$p = 1/1 + exp(-(\beta 0 + \beta 1\, X + \beta 2\, X2 + \ldots + \beta p\, Xp))$$

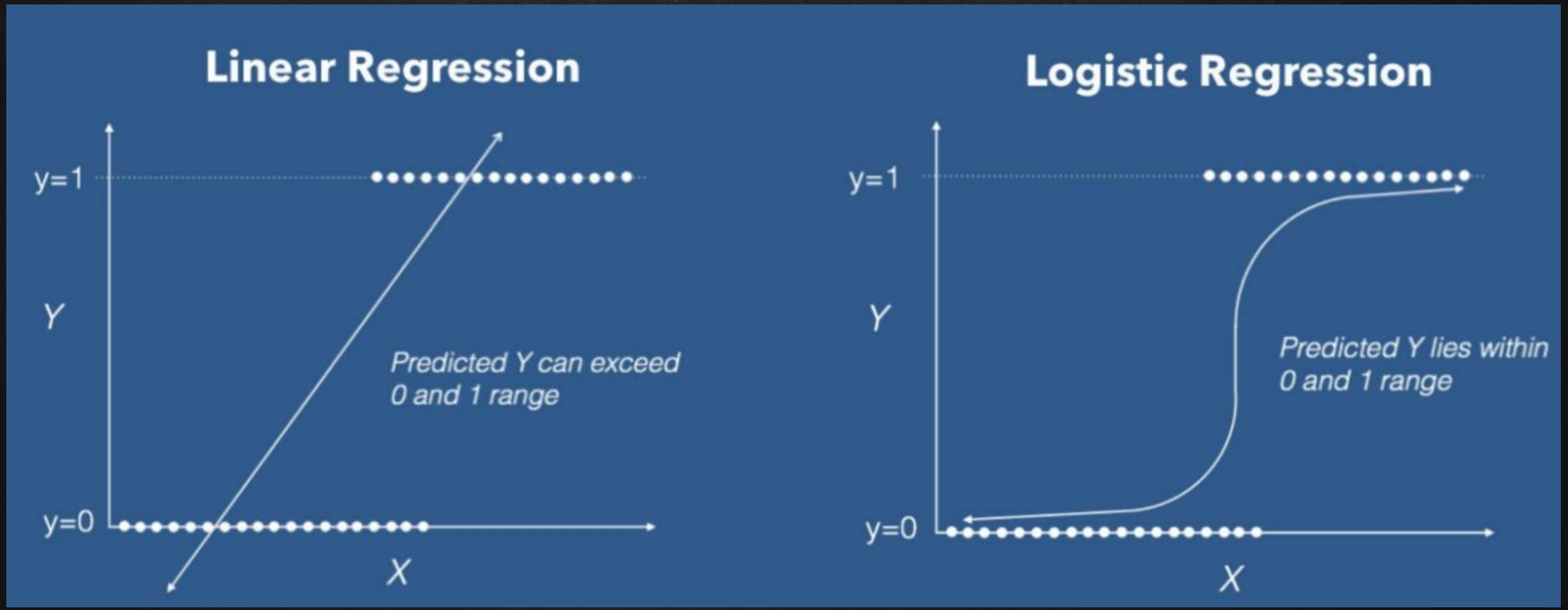
$$Odds(Y=1) = p/1-p, \text{ therefore } p = Odds/1+Odds$$

$$Odds(Y=1) = exp(\beta 0 + \beta 1\, X + \beta 2\, X2 + \ldots + \beta p\, Xp)$$

This is the linear function:

$$\log(Odds(Y=1)) = \beta 0 + \beta 1\, X + \beta 2\, X2 + \ldots + \beta p\, Xp$$

# Linear regression VS logistic regression



# Cost Function

$$J(\theta) = -\frac{1}{m}\sum\left[y^{(i)}\log(h\theta(x(i))) + \left(1-y^{(i)}\right)\log(1-h\theta(x(i)))\right]$$

Here,
hθ(x) = hβ(x) = ρ = 1/1 + exρ(–(β0 + β1 X + β2 X2 + … + βρ Xρ))

Sigmoid function:

$$f(x) = \frac{1}{1+e^{-(x)}}$$

MoreReading

# Evaluation of classification model

**Confusion matrix**: “A tabular display (2×2 in the binary case) of the record counts by their predicted and actual classification status”.

**Sensitivity**: “The percent (or proportion) of all 1s that are correctly classified as 1s. Also called Recall “
**Specificity**: “The percent (or proportion) of all 0s that are correctly classified as 0s.”
**Precision**: “The percent (proportion) of predicted 1s that are actually 1s.”

**ROC curve**: “A plot of sensitivity versus specificity.”
**AUC:** “Area under a ROC curve”

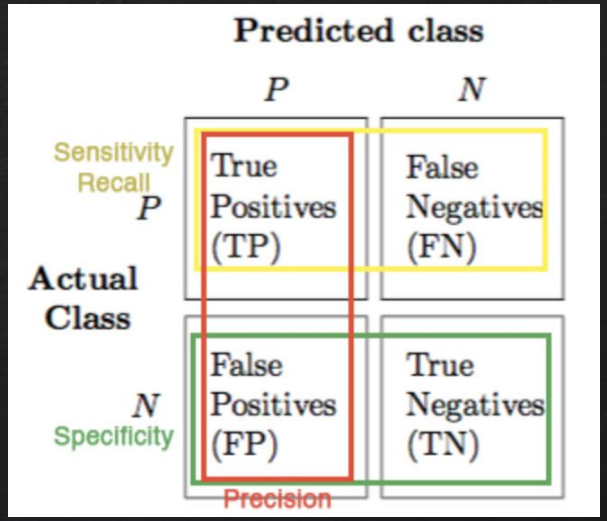
**Lift**: “A measure of how effective the model is at identifying (comparatively rare) 1s at different probability cutoffs.”

**Accuracy**: ”The percent (or proportion) of cases classified correctly.”

Other popular classification model evaluation metrics:
F1, KS, SomersD, GINI coefficient

# Confusion matrix and key concepts

**Confusion matrix**: “A tabular display (2×2 in the binary case) of the record counts by their predicted and actual classification status”.



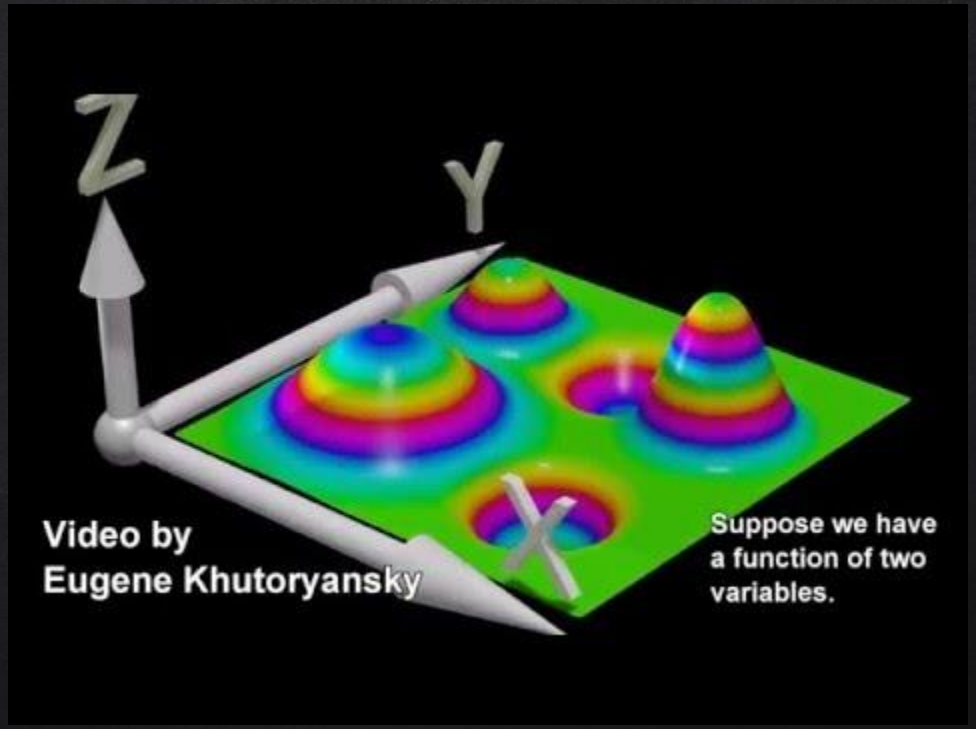

[Classification Evalu. Reading]


# Gradient descent

This algorithm is the most popular and one fit all method used in Machine learning to searching for the solution which minimize the cost function. This is a way to find answer other than solving a matrix equation

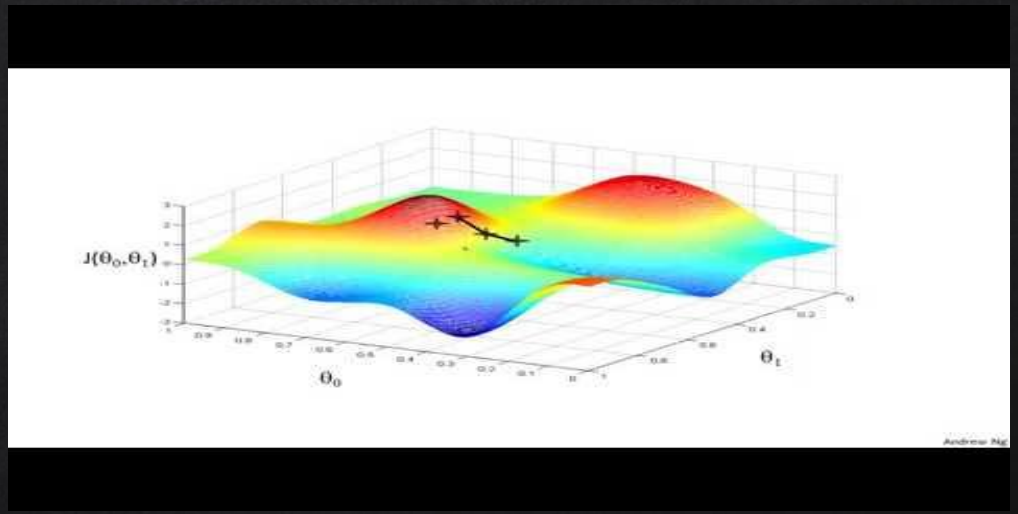
The key concepts are the cost function, (partial) derivative, learning rate

# Gradient and Partial derivative




From Physics Videos by Eugene Khutoryansky

# Gradient descent explained/implementation by Andrew Ng



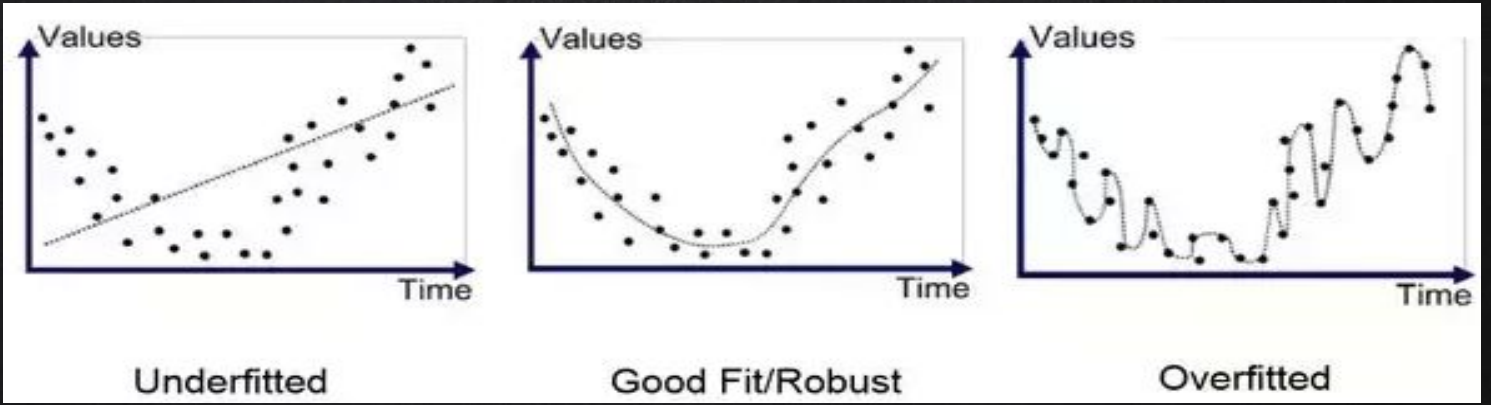

For whom have interested to learn more


# Regularization

"In mathematics, statistics, and computer science, particularly in machine learning and inverse problems, regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting.[1]"

– Wiki

# REGULARIZATION – OVERFITTING CONTROL



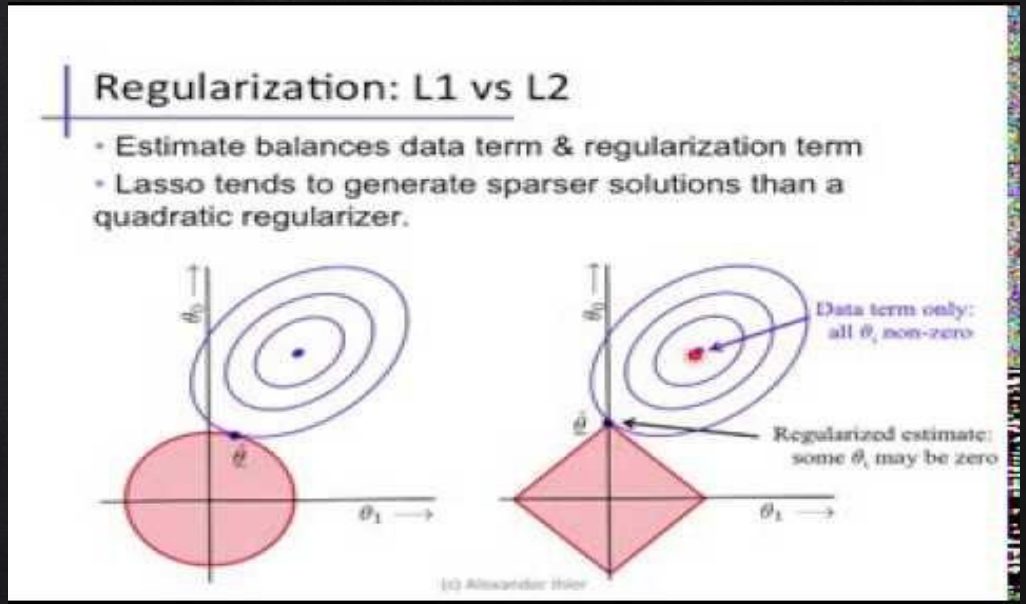

Three types of regularization in linear models, add regularization(penalized) term to cost function

Lasso regression (L1) : New Cost Function = Cost Function + $\lambda \sum_{j=1}^{p} |\beta_j|$

Ridge regression (L2) : New Cost Function = Cost Function + $\lambda \sum_{j=1}^{p} \beta_j^2$

Elastic Net regression (L1+L2) : New Cost Function = Cost Function + $\lambda \sum_{j=1}^{p} |\beta_j|$ + $\lambda \sum_{j=1}^{p} \beta_j^2$

# L1 and L2 regularization explained




From Alexander Ihler channel

# Recommended Readings

https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3

https://newbiettn.github.io/2016/08/30/precision-recall-sensitivity-specificity/

https://towardsdatascience.com/regularization-an-important-concept-in-machine-learning-5891628907ea

https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

Shared book and paper in github repo

# References


https://www.edureka.co/blog/statistics-and-probability/#Inferential%20Statistics

https://newbiettn.github.io/2016/08/30/precision-recall-sensitivity-specificity/

https://www.autodeskresearch.com/publications/samestats

https://www.edureka.co/blog/supervised-learning/#overview

https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

https://online.stat.psu.edu/stat462/node/93/

Bruce, Peter,Bruce, Andrew,Gedeck, Peter. Practical Statistics for Data Scientists (Kindle Location 5267). O'Reilly Media. Kindle Edition.

Physics Videos by Eugene Khutoryansky

Stanford university , Machine Learning by Andrew Ng

https://en.wikipedia.org/wiki/Regularization_(mathematics)

https://www.quora.com/What-is-regularization-in-machine-learning