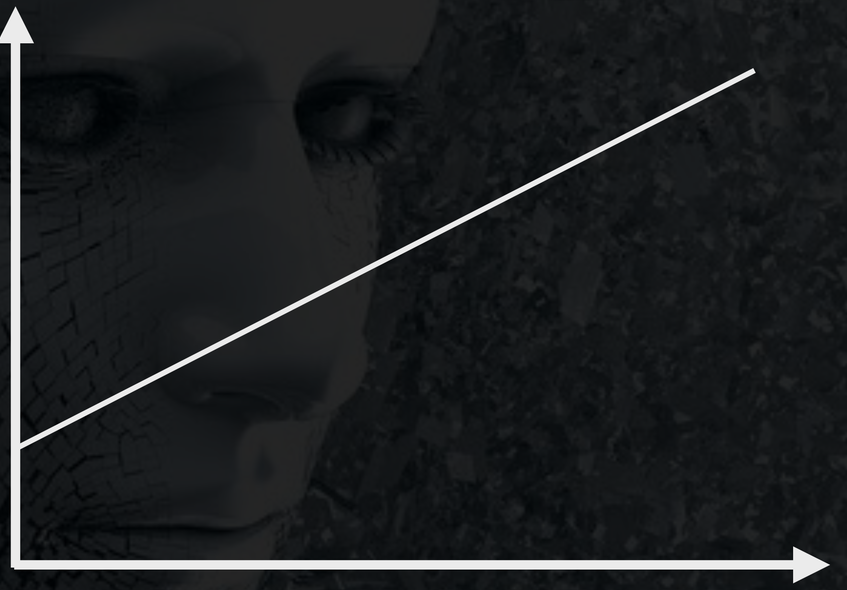
LINEAR REGRESSION

LINEAR REGRESSION

Cause

Effect

- Regression Analysis is one of the most common method of making predictions & its based out cause & effect relationship between variables.
- Income is the function of education or good grades are the function of time you studied



Xilytica

LINEAR REGRESSION

SAT	GPA
1714	2.40
1664	2.52
1760	2.54
1685	2.74
1693	2.83
1670	2.91
1764	3.00
1764	3.00
1792	3.01

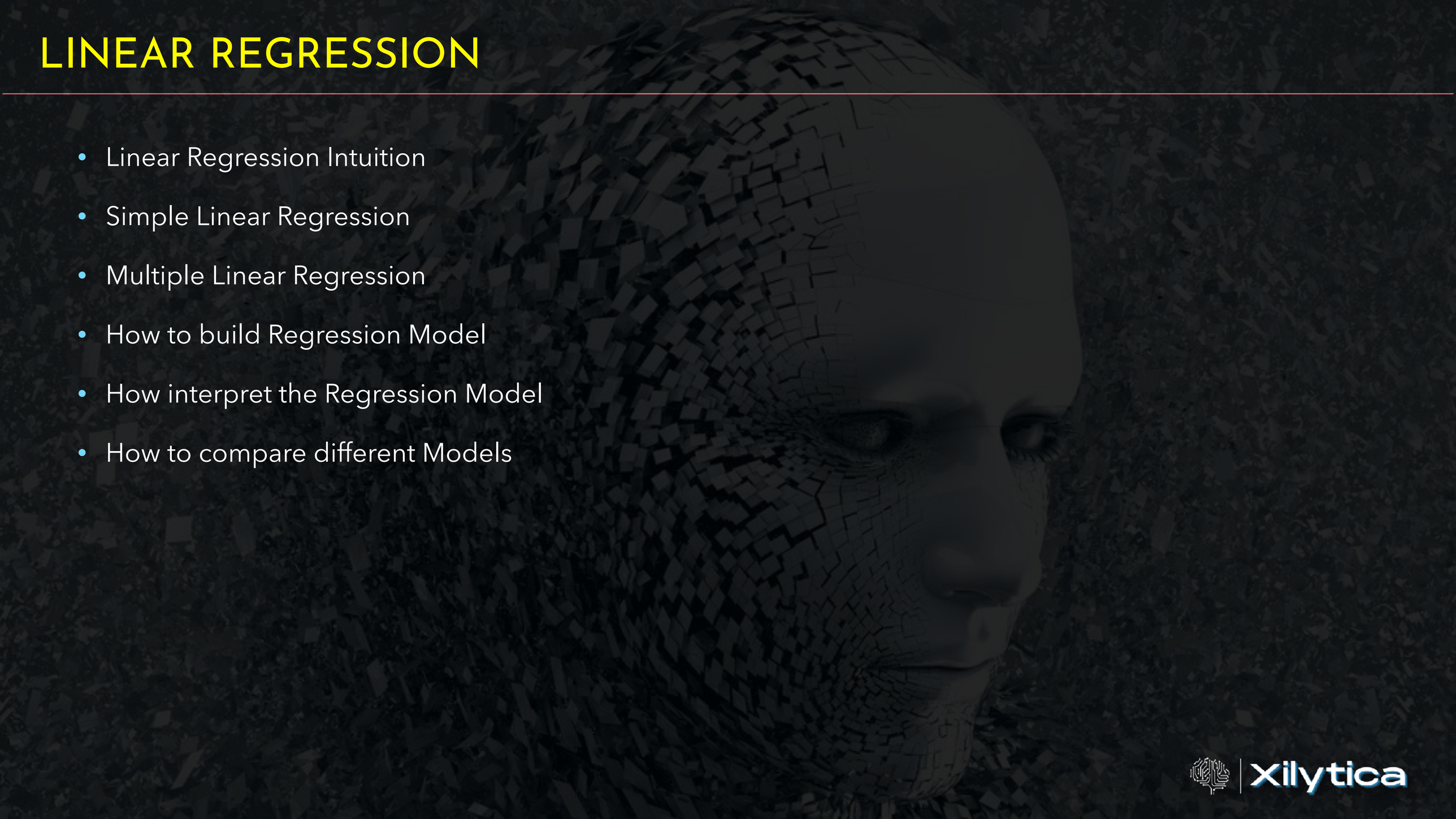
Experience	Salary
1.1	39343.0
1.3	46205.0
1.5	37731.0
2.0	43525.0
2.2	39891.0
2.9	56642.0

Brand	Price	Body	Mileage	EngineV	Engine Type	Registration	Year	Model
BMW	4200.0	sedan	277	2.0	Petrol	yes	1991	320
Mercedes-Benz	7900.0	van	427	2.9	Diesel	yes	1999	Sprinter 212
Mercedes-Benz	13300.0	sedan	358	5.0	Gas	yes	2003	S 500
Audi	23000.0	crossover	240	4.2	Petrol	yes	2007	Q7
Toyota	18300.0	crossover	120	2.0	Petrol	yes	2011	Rav 4

- Linear Regression describes the relationship between two or more variables. In Simple Linear regression we have only two variables independent variable & dependent variable. the dependent variable must be continuous on the other hand out independent variables can be categorical or numerical.
- In Multiple Linear Regression we have one dependent variable and more than one independent variables.

LINEAR REGRESSION

- Linear Regression Intuition
- Simple Linear Regression
- Multiple Linear Regression
- How to build Regression Model
- How interpret the Regression Model
- How to compare different Models



Xilytica

LINEAR REGRESSION

Linear Regression : It is the linear approximation of a causal relationship between two or more variables. Linear model are highly popular for making inferences & predictions

Independent Variables / Predictors

Dependent Variables / Predicted

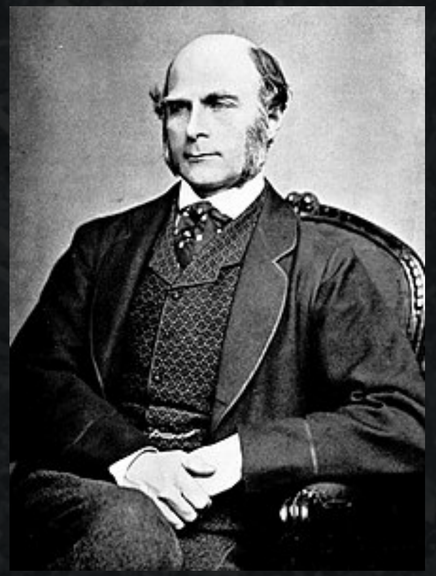
$$Y = F(x1,x2,x3...xk)$$

The dependent variable Y is the function of independent variables x1 to xk

Xilytica

LINEAR REGRESSION

- In 1800s Francis Galton was studying relationship between parents and their children. In particular, he investigated the relationship between height of fathers and their sons.
- He discovered that sons hight tends to be roughly as tall as his father.
- However, Galton's breakthrough was that the son's height tended to be closer to the overall average height of all people.
- Shaquille O'Neal - height 7ft 1 Very all almost an outlier or anomaly.



LINEAR REGRESSION

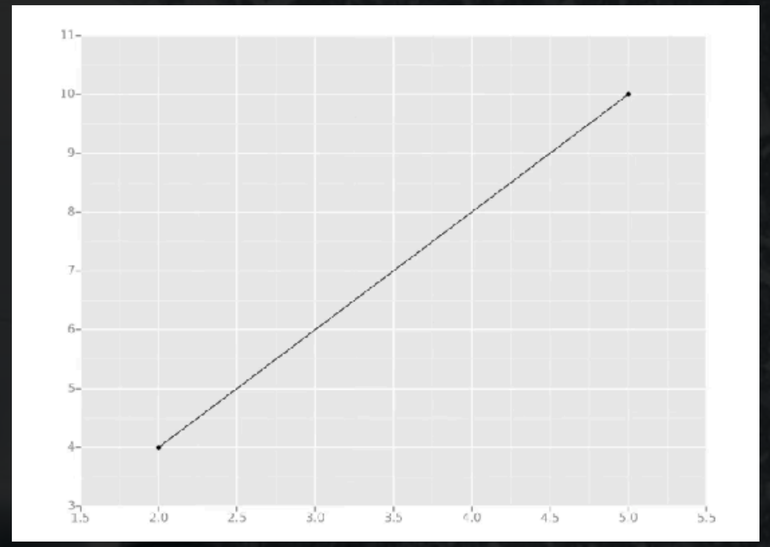
- Shaquille O'Neal - height 7ft 1. Very Tall all almost an outlier or anomaly.
- Shaq's son is also very tall but not as tall as his dad.
- Galton call this phenomenon as regression.
- As a father's son's height tends to regress or drift towards the mean or average height.

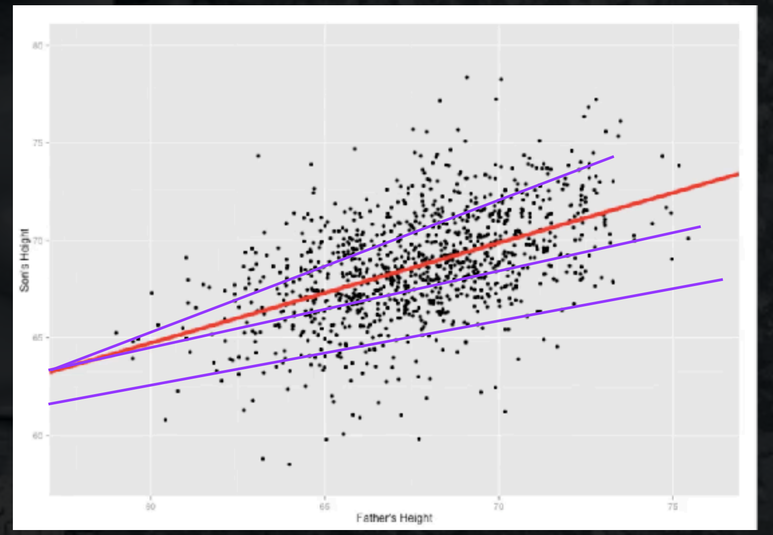
Xilytica

LINEAR REGRESSION

- Calculating regression with only two points.
- We try to draw a regression line such that it is as close to all data points as possible.
- In Classic Linear Regression or Least Square Method you only measure the distance in up wards and down wards direction.

Our goal here is to determine the best line that has minimum vertical distance between all data points and our line.

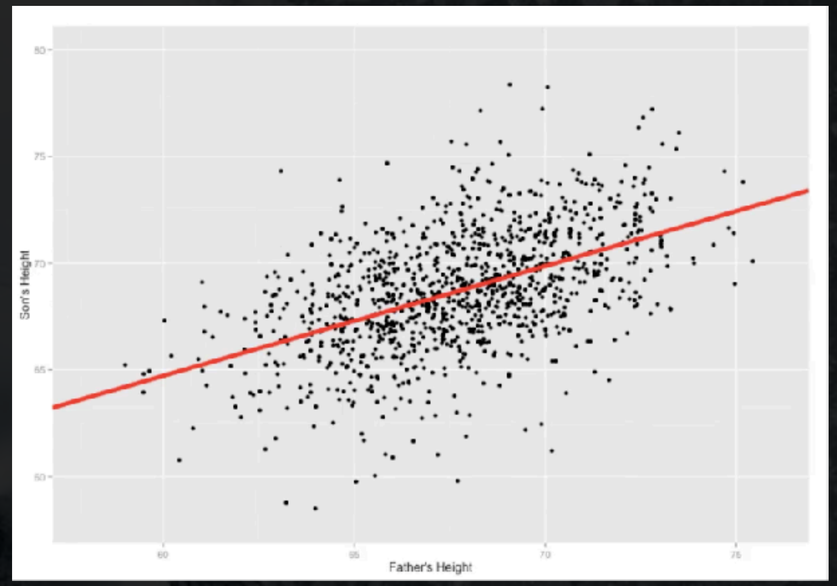




Xilytica

LINEAR REGRESSION

- On X axis we have Father’s height and on Y axis we have Son’s height. Here we have mapped multiple fathers' & son’s heights.
- Using regression line we can predict son’s height even before he is born.



Xilytica

SIMPLE LINEAR REGRESSION

Experience	Salary
1.1	39343.0
1.3	46205.0
1.5	37731.0
2.0	43525.0
2.2	39891.0
2.9	56642.0

Simple
Linear Regression

Constant

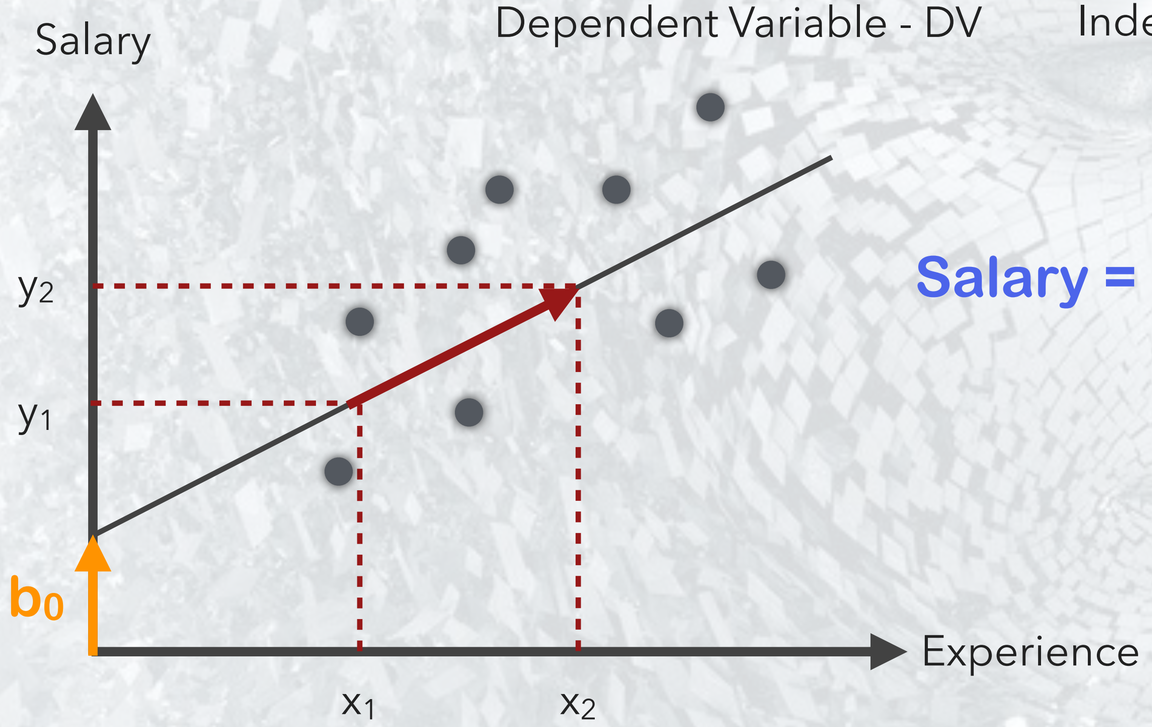
Coefficient - Quantify the effects of x on y

$$y = b_0 + b_1 * x + \varepsilon$$

Dependent Variable - DV

Independent Variable - IDV

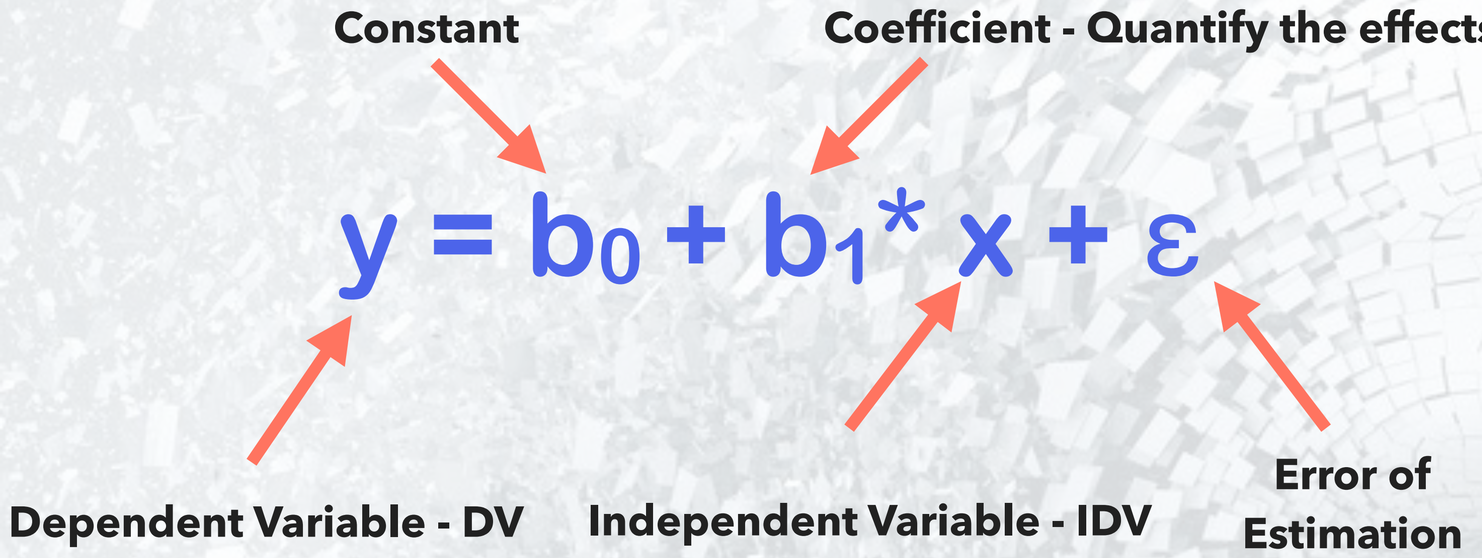
Error of Estimation



Salary = b_0 + b_1* Experience

Xilytica

SIMPLE LINEAR REGRESSION



$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

b0 - is the 1st coefficient that we will try to learn. This is the salary of a person if he has no experience that is experience = 0

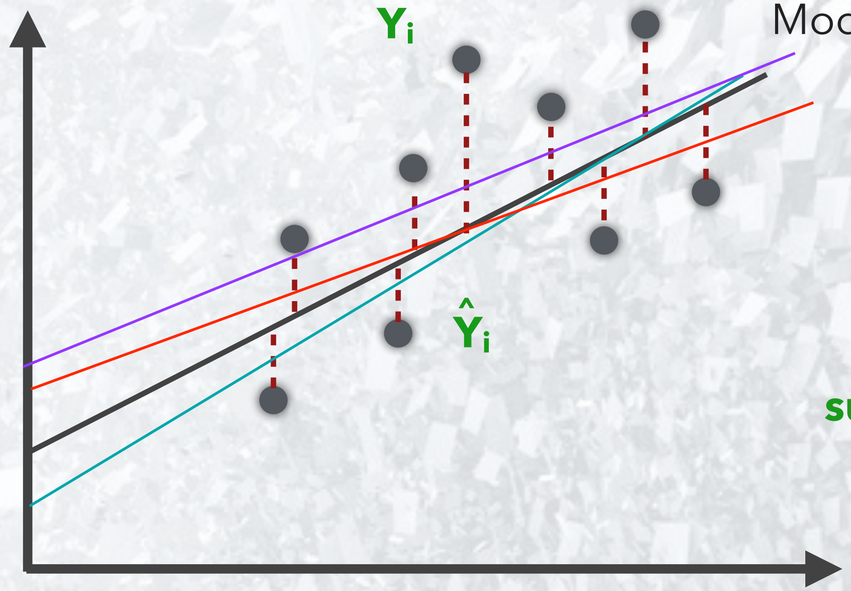
b1 - is the 2nd coefficient that is the slope of the line. It means for each year of experience how much the salary would increase

X - Its the total year of experience

y - Salary - this is what we want to predict

Xilytica

SIMPLE LINEAR REGRESSION



Y_i - Observed Value

$\hat{Y}_i$ - Predicted Value

$Y_i - \hat{Y}_i$ - Observed Value - Expected Value

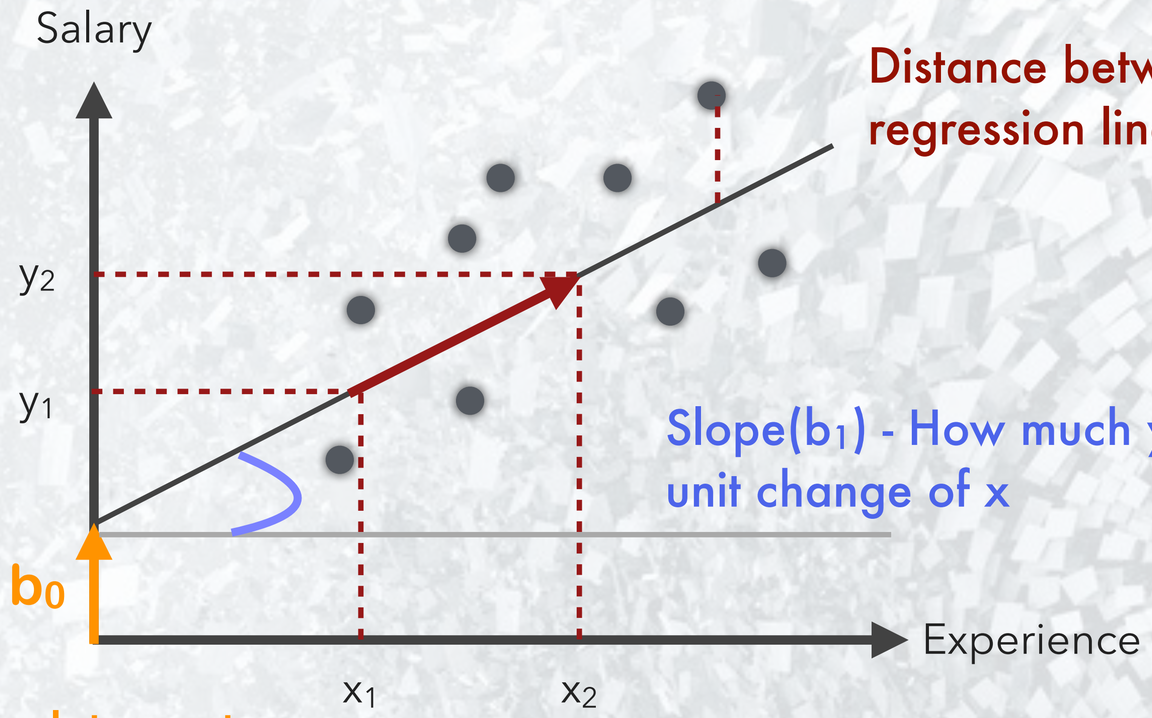
$sum(Y_i - \hat{Y}_i)^2$ - Minimum

$$Salary = b_0 + b_1 * Experience$$

So we try to find the best fit line using the b0 & b1 values. That is the line that has the least cost function.

Xilytica

SIMPLE LINEAR REGRESSION



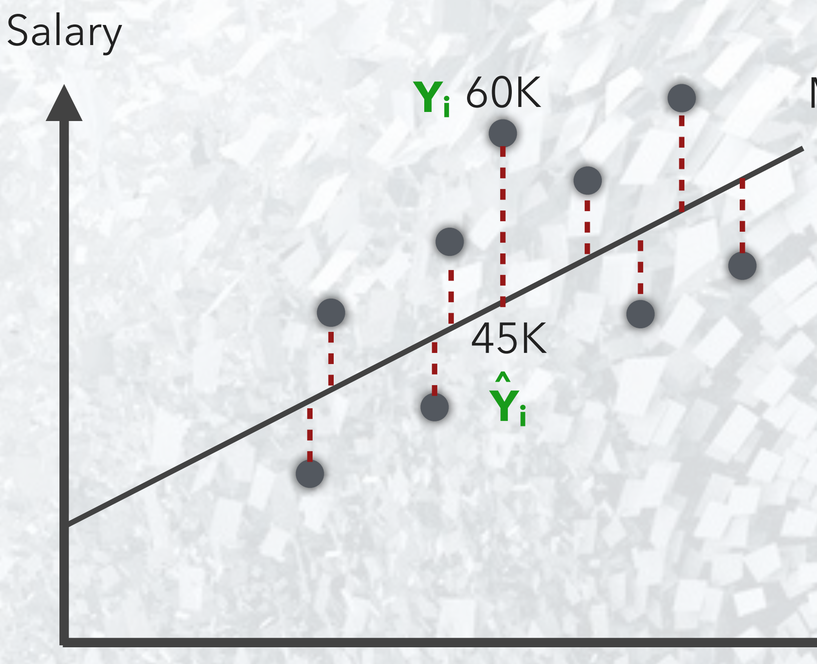
Distance between the observed value & regression line is the estimation of the error

Slope(b_1) - How much y changes for each unit change of x

b_0 - Intercept

$$y = b_0 + b_1 * x + \varepsilon$$

SIMPLE LINEAR REGRESSION



$$\text{Salary} = b_0 + b_1 * \text{Experience}$$

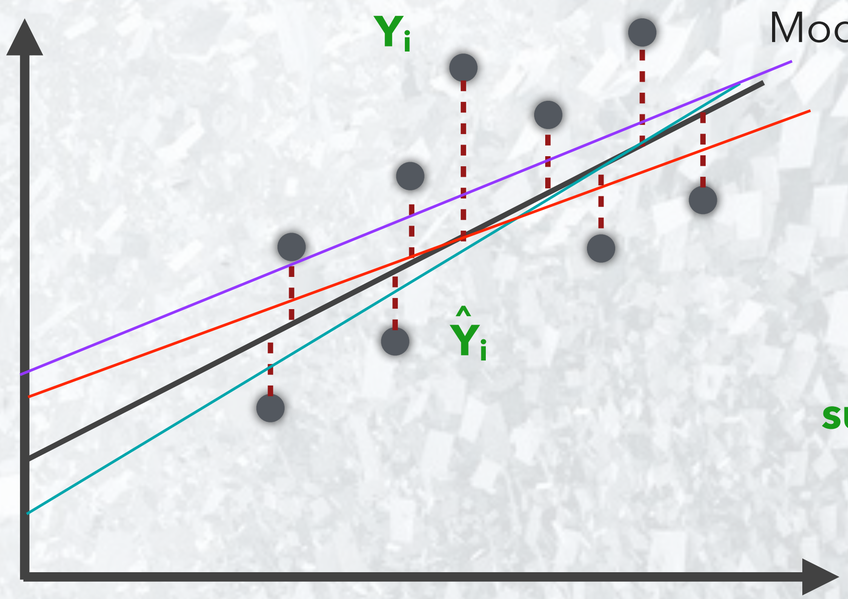
Y_i - Salary person is actually getting

$\hat{Y}_i$ - Salary person should get as per model line

$Y_i - \hat{Y}_i$ - Observed Value - Expected Value

Xilytica

BEST FITLINE



Y_i - Observed Value

$\hat{Y}_i$ - Modelled Value

$Y_i - \hat{Y}_i$ - Observed Value - Expected Value

$sum(Y_i - \hat{Y}_i)^2$ - Minimum

Linear Regression draws possible trends line and count the minimum vertical distance between Observed Value and the Trend Line.

$sum(Y_i - \hat{Y}_i)^2$ -> Minimum

Xilytica

OLS - Ordinary Least Squares

OLS is a method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the sum of the squares of the differences between the target dependent variable and those predicted by the linear function. In other words, it tries to minimizes the sum of squared errors (SSE) or mean squared error (MSE) between the target variable (y) and our predicted output ($\hat{y}$) over all samples in the dataset.

OLS can find the best parameters using of the following methods:

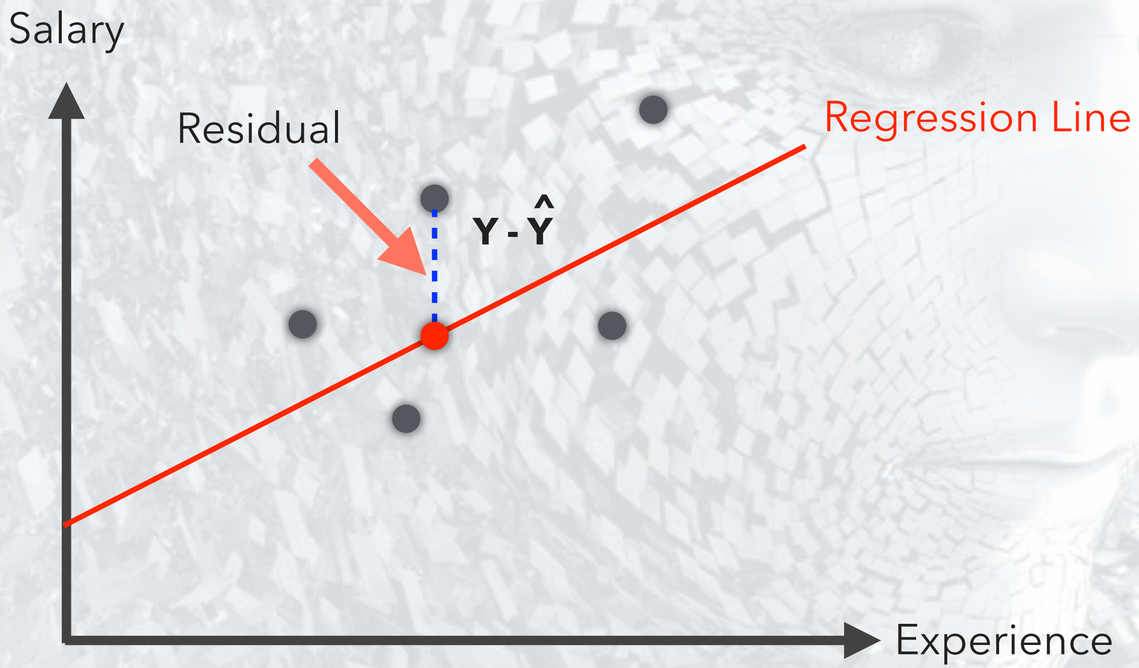
- Solving the model parameters analytically using closed-form equations
- Using an optimization algorithm (Gradient Descent, Stochastic Gradient Descent, Newton's Method, etc.)

LINEAR REGRESSION

Residuals Residuals are the difference between actuals values of the variables you are predicting and predicted values from your model.

$$e = Y - \hat{Y}$$

For most regressions you want residuals to be normally distributed. That implies the sum & mean of the difference between actual values and predicted values is close to zero.



Xilytica

LINEAR REGRESSION

The objective of linear regression is to minimize the cost function

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$h_\theta(x) = \theta^T x = \theta_0 + \theta_1 x_1$$

$\theta^T X$ is our vector form of a line equation

h_θ (x) is the predicted value

y - is the actual value

$$y = b_0 + b_1 * x$$

Xilytica

LINEAR REGRESSION

erro = (b0 +b1*x) - Y

$$ypred = b_0 + b_1 * x$$

$$error = (b_0 + b_1 * x) - y$$

$$\text{Mean Square error} = 1/m \sum ((b_0 + b_1 * x) - y)^2$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

As this error can be negative & the other value could be positive will reduce or cancel the the error. We are more interested in getting the magnitude of error. So either will take the absolute values or take the Euclidean distance between the lines. That is the actual value & the predicted value.

Xilytica

MSE - Mean Square Error

The objective of linear regression is to minimize the cost function

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

We try to minimise this mean square error & this will give us the best fit line. This is our cost function which is we want to minimise. We have divided by 2 & for ease of calculation we calculating this using calculus

SSE - Sum of Square Error

$$SS_{(residuals)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

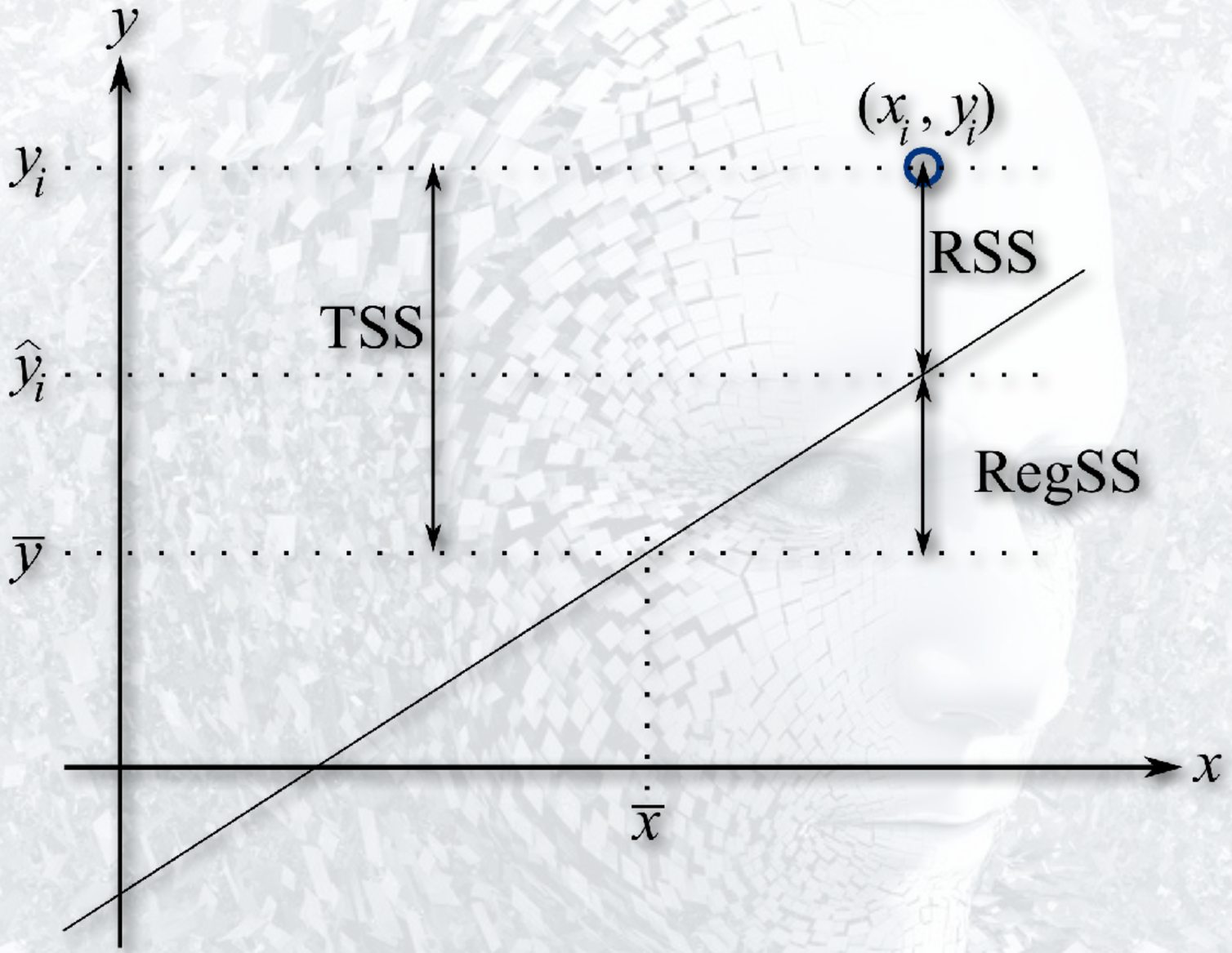
Sum of Squared error is an error that we can't explain in our model.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

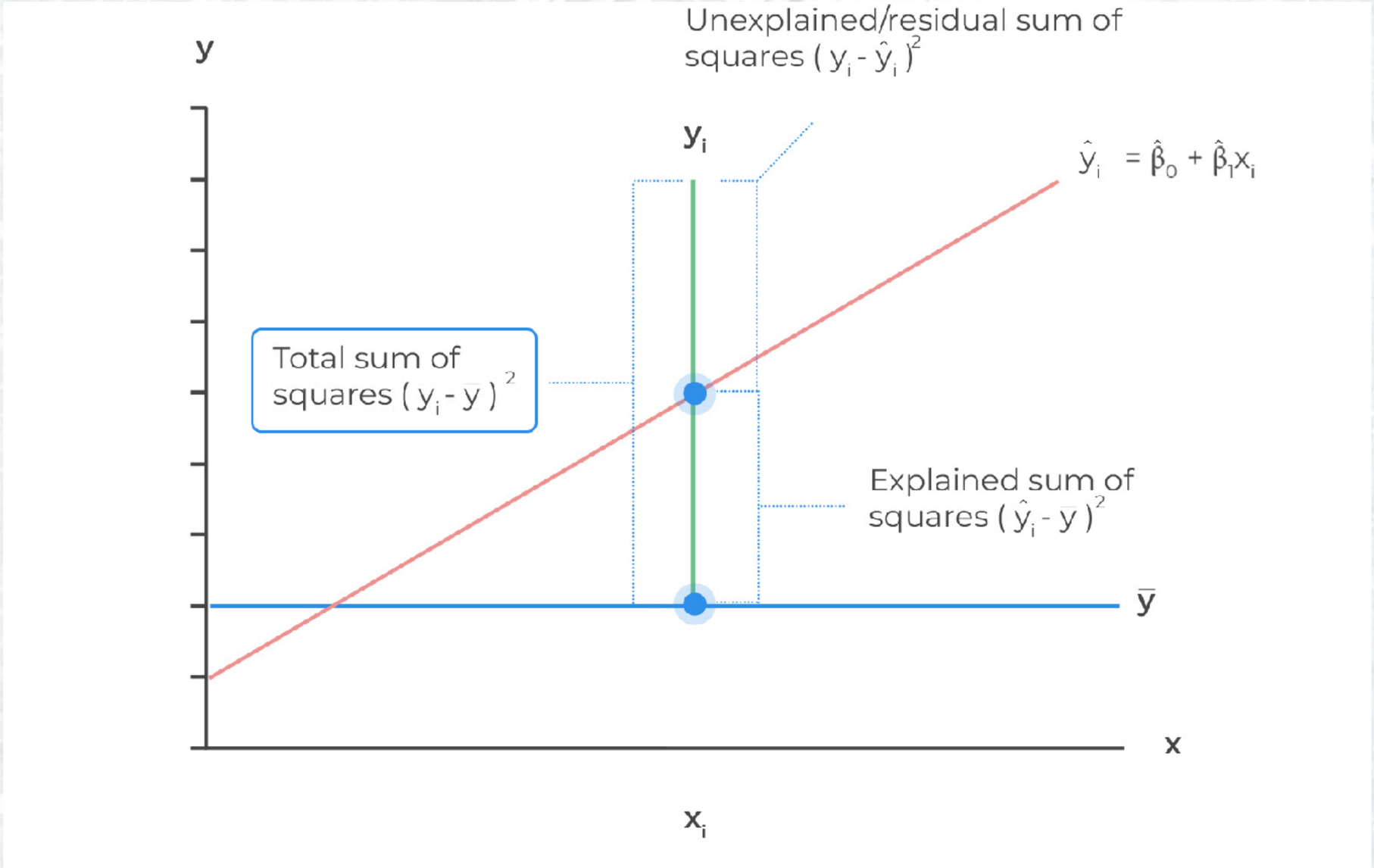
Total Sum of Squares - Measures the distance between our actual value & the actual average value. Total squared error is the total variation.

Xilytica

TSS - Sum Squares



TSS - Sum Squares



R Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R Squared - It is the measure of the explained variation by our model

R Squared = 1 - Unexplained Variance/ Total Variance

R Squared = Explained Variance/Total Variance

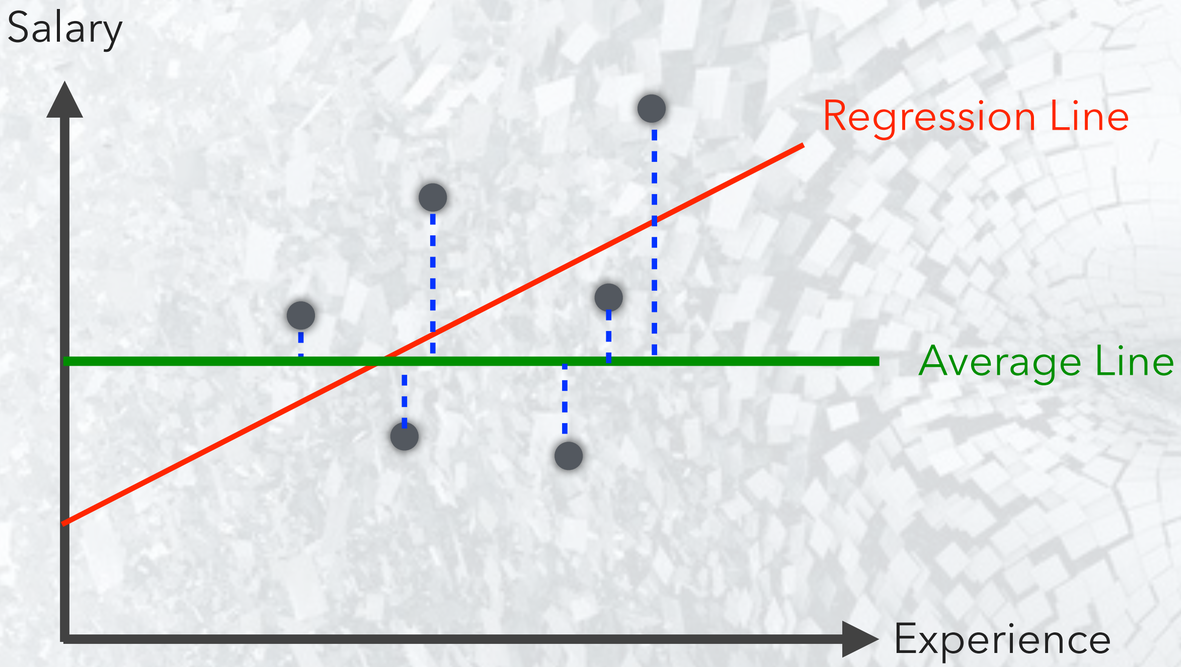
It means how well our model explains the variation from the mean. That is how much are we able to reduce the amount of the unexpected variance. Closer it is to the 1 better it is regards to the explaining the overall variance.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

SST = SSR + SSE

Total Variability of Dataset = Explained Variability by Regression Line + Unexplained Variability know as error

R Squared



$$SS_{res} = \text{sum}(Y_i - \hat{Y}_i)^2$$

$$SSt_{ot} = \text{sum}(Y_i - \hat{Y}_{avg})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Trying to fit a line to minimise SS_{res} as small as possible. R2 tells how good is your regression line compared to the average line. As sum of square of residuals is zero
R2 is 1. R2 will increase by adding new variables.

Xilytica

Regression Evaluation Matrix

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$R^2 = 1 - RSE$$

MAE - Mean of the absolute error. Average of the errors.

MSE - Mean Square Error

RMSE - Root Mean Squared Error

RAE - Relative Absolute Error

RSE - Relative Squared Error

R^2 - R Squared is the not the error but popular matrix to evaluate the matrix. It measure how closer the values are to the line. Higher the R squared better the model fits the data.

R Squared

$$RAE = \frac{\sum_{j=1}^{n} |y_j - \hat{y}_j|}{\sum_{j=1}^{n} |y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2}$$

$$R^2 = 1 - RSE$$

- The R Squared - Measures how much of the total variablility is explained by the model.
- R2 also tells you how good is your regression line compared to your average line.
- R2 will increase by adding a new variables.
- Multiple regressions are always better than simple linear regression as with each additional variable you add, the
- explainatory power may only increase or remain the same

ADJUSTED R SQUARE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

p - Number of regressors or number of features
n - Sample Size or No. of rows

Adjusted R square has a penalising factor that means it penalise for adding a new independent variable that does not help your model.

Mean Square error is one such error metric for judging the accuracy and error rate of any machine learning algorithm for a regression problem.
So, MSE is a risk function that helps us determine the average squared difference between the predicted and the actual value of a feature or variable.

ACCURACY OF A MODEL

- Adjusted R square has a penalising factor that means it penalise for adding a new independent variable that does not help your model.

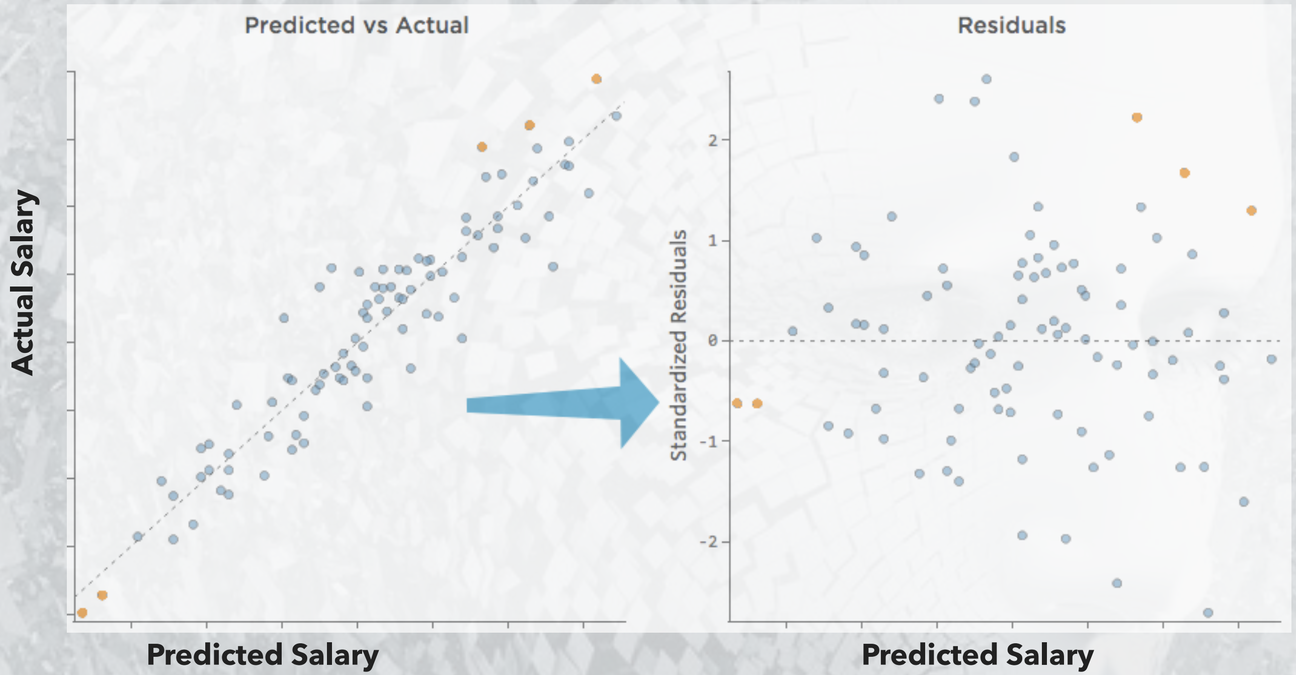
- Mean Square error is one such error metric for judging the accuracy and error rate of any machine learning algorithm for a regression problem.
- So, MSE is a risk function that helps us determine the average squared difference between the predicted and the actual value of a feature or variable.

- RMSE is an acronym for Root Mean Square Error, which is the square root of value obtained from Mean Square Error function.
- Using RMSE, we can easily plot a difference between the estimated and actual values of a parameter of the model.
- By this, we can clearly judge the efficiency of the model.

- Usually, a RMSE score of less than 180 is considered a good score for a moderately or well working algorithm. In case, the RMSE value exceeds 180, we need to perform feature selection and hyper parameter tuning on the parameters of the model.
-

INTERPRETING OUR MODEL

Residuals The distance from the line at zero shows how bad the predicted values is.



http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/

Xilytica

INTERPRETING OUR MODEL

Sum Of Square Errors	Measures the unexplained variability by the regression
Standard Error of the Coefficient Estimate	Measure of the variability in the estimate for the coefficient. Lower means better but this number is relative to the value of the coefficient. This value to be at least an order of magnitude less than the coefficient estimate.
t-value of the Coefficient Estimate	Is used to calculate the p-value and the significance levels.
Variable p Value	Probability the variable is *NOT* relevant. You want this number to be as small as possible.
R-squared	Metric for evaluating the goodness of fit of your model. Higher is better with 1 being the best. Corresponds with the amount of variability in what you're predicting that is explained by the model. **WARNING**: While a high R-squared indicates good correlation, correlation does not always imply causation.

Xilytica

INTERPRETING OUR MODEL

Residual Std Error / Degrees of Freedom

The Residual Std Error is just the standard deviation of your residuals. You'd like this number to be proportional to the quantiles of the residuals in #1. For a normal distribution, the 1st and 3rd quantiles should be 1.5 +/- the std error.

The Degrees of Freedom is the difference between the number of observations included in your training sample and the number of variables used in your model (intercept counts as a variable).

F-statistic & resulting p-value

Performs an F-test on the model. This takes the parameters of our model (in our case we only have 1) and compares it to a model that has fewer parameters. In theory the model with more parameters should fit better. If the model with more parameters (your model) doesn't perform better than the model with fewer parameters, the F-test will have a high p-value (probability NOT significant boost). If the model with more parameters is better than the model with fewer parameters, you will have a lower p-value.

The DF, or degrees of freedom, pertains to how many variables are in the model. In our case there is one variable so there is one degree of freedom.

CORRELATION

The term correlation is a combination of two words ‘Co’ (together) and relation (connection) between two quantities. Correlation is when, at the time of study of two variables, it is observed that a unit change in one variable is retaliated by an equivalent change in another variable, i.e. direct or indirect.

cor (x,y)

FORM	DIRECTION	STRENGTH	OUTLIERS
Linear, Quadratic, Non-Linear	Positive, Negative	Scattered, Strong	Outliers

Correlation Vs Regression

CORRELATION

Determines the degrees of relationship between two variables

Used to represent linear relationship between two variables. Relationship between f(x, y) is same as f(y, x)

Indicates the strength of association between variables.

Aims at finding a numerical value that expresses the relationship between variables.

REGRESSION

How one variable affects other or describes how an independent variable is related to dependent variable.

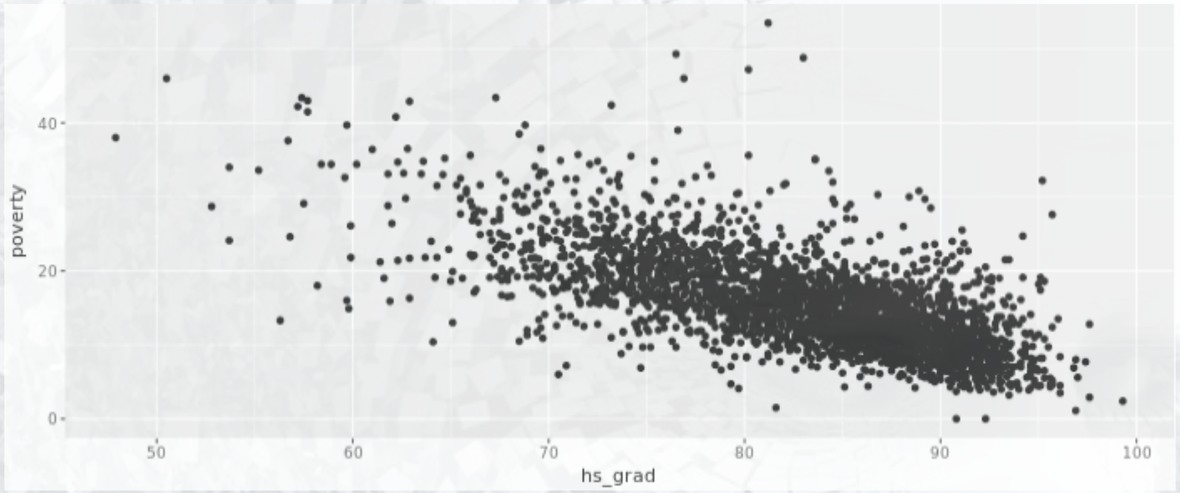
Used to fit the best line and estimate one variable on the basis of another. Its one way.

Reflects the impact of the unit change in independent variable on the dependent variable.

Goal is to predict value of the random variable on the basis of the values of the fixed variable.

Correlation Vs Regression

The correlation coefficient between the poverty rate of counties in the United States and the high school graduation rate in those counties was -0.681. Choose the correct interpretation of this value.

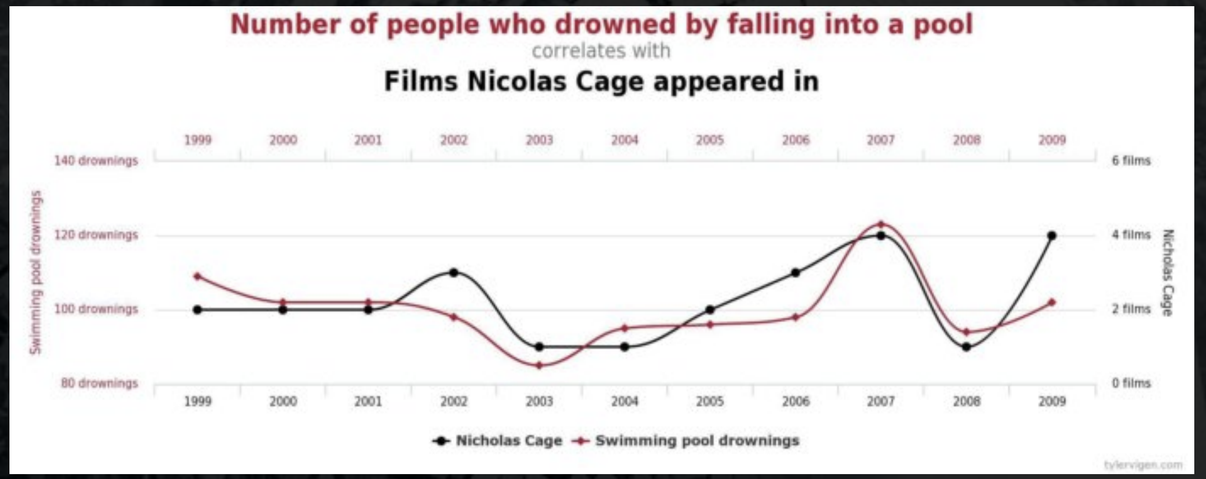


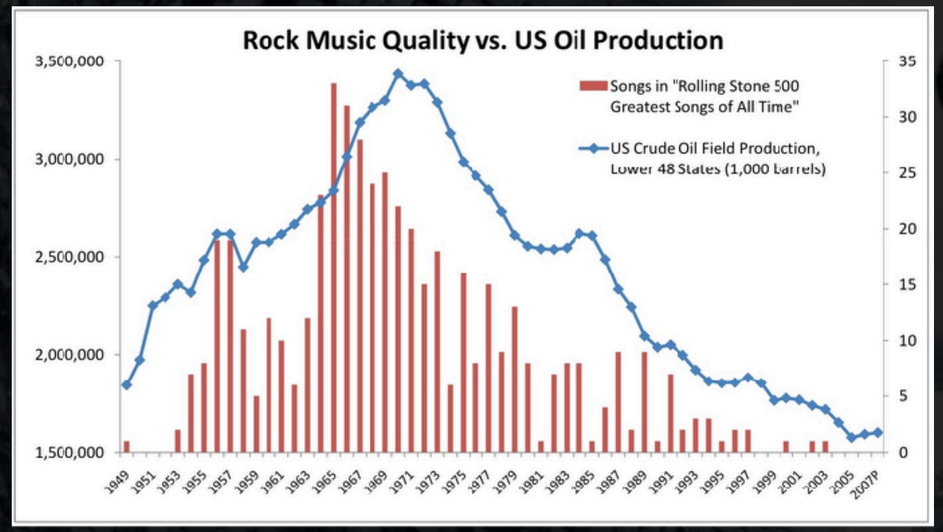
- ◯ Having a higher percentage of high school graduates in a county results in that county having lower poverty rates.
- ◯ Because the correlation is negative, there is no relationship between poverty rates and high school graduate rates.
- ◯ Counties with lower high school graduation rates are likely to have higher poverty rates.

Xilytica

Correlation

CORRELATION DOES NOT MEAN CAUSATION



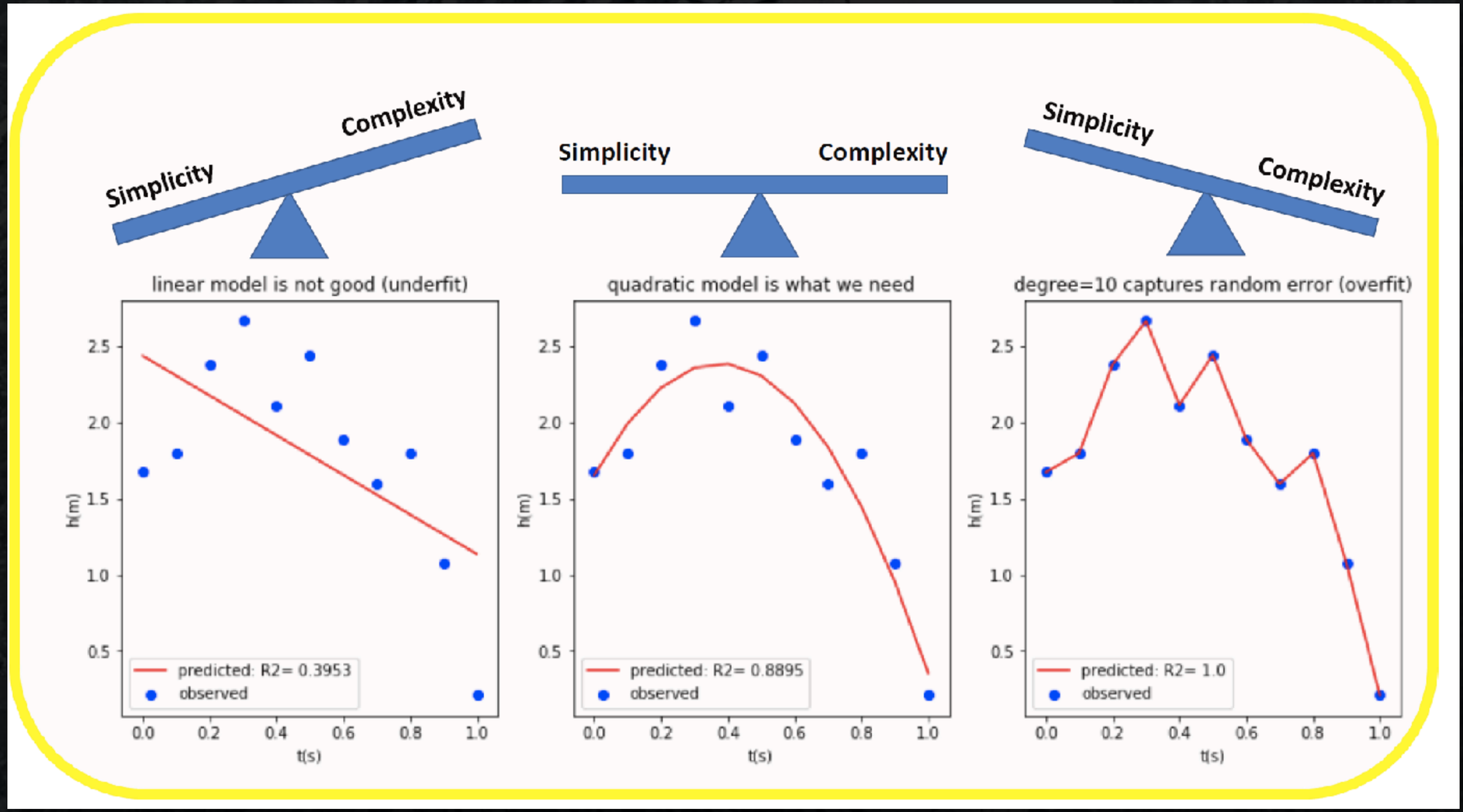


http://phenomena.nationalgeographic.com/2015/09/11/nick-cage-movies-vs-drownings-and-more-strange-but-spurious-correlations/

https://www.overthinkingit.com/2008/09/23/the-hubbert-peak-theory-of-rock-or-why-were-all-out-of-good-songs/

Xilytica

Simple Vs Complex Models



Xilytica

Simple Vs Complex Models

Model complexity can be characterized by many things, and is a bit subjective.

In machine learning, model complexity often refers to the number of features or terms included in a given predictive model, as well as whether the chosen model is linear, nonlinear, and so on. It can also refer to the algorithmic learning complexity or computational complexity.

Overly complex models are less easily interpreted, at greater risk of overfitting, and will likely be more computationally expensive. There are some really sophisticated and automated methods by which to control, and ultimately reduce model complexity, as well as help prevent overfitting. Some of them are able to help with feature and model selection as well.

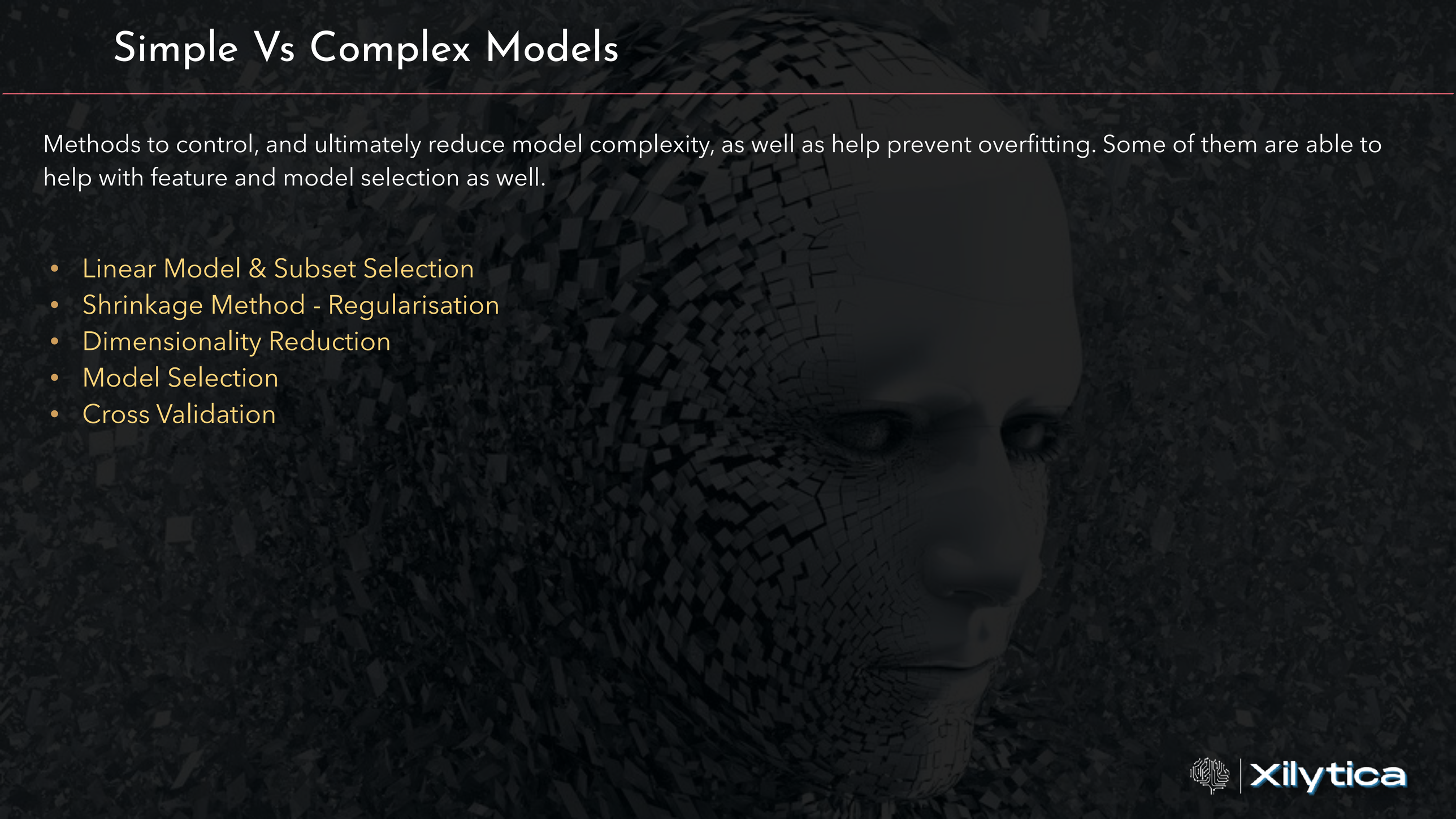
These methods include linear model and subset selection, shrinkage methods (including regularization), and dimensionality reduction.

https://www.innoarchitech.com/blog/machine-learning-an-in-depth-non-technical-guide-part-3

Simple Vs Complex Models

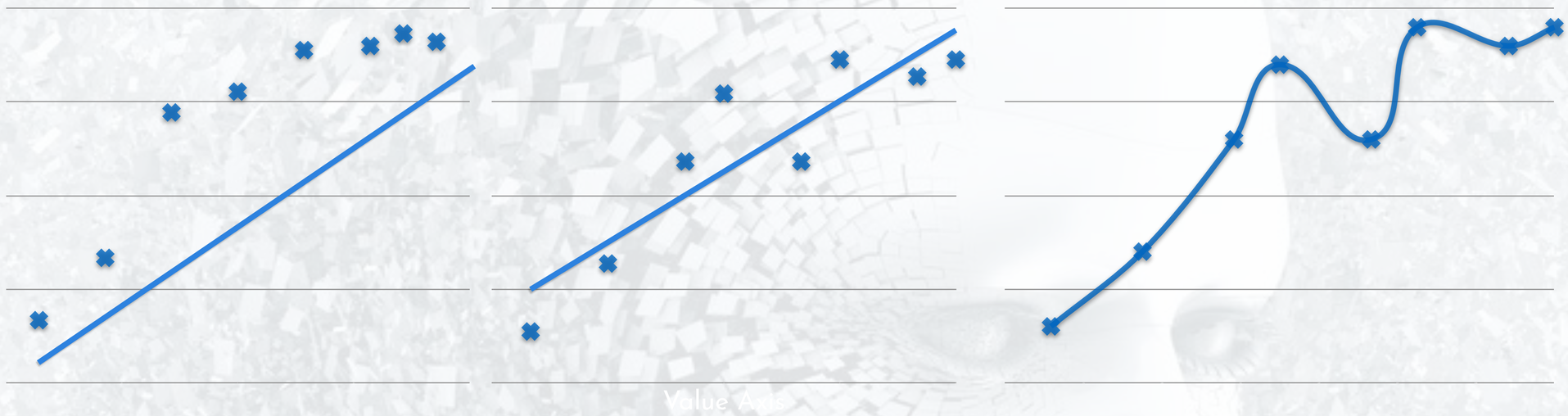
Methods to control, and ultimately reduce model complexity, as well as help prevent overfitting. Some of them are able to help with feature and model selection as well.

- Linear Model & Subset Selection
- Shrinkage Method - Regularisation
- Dimensionality Reduction
- Model Selection
- Cross Validation



Model Fitness

Model Fitness



Under-fitting

High error from training data points will also have high error on test data. ie Not perform good on training & test data

Good-fitting

Low training error. Good fit for both training & test data

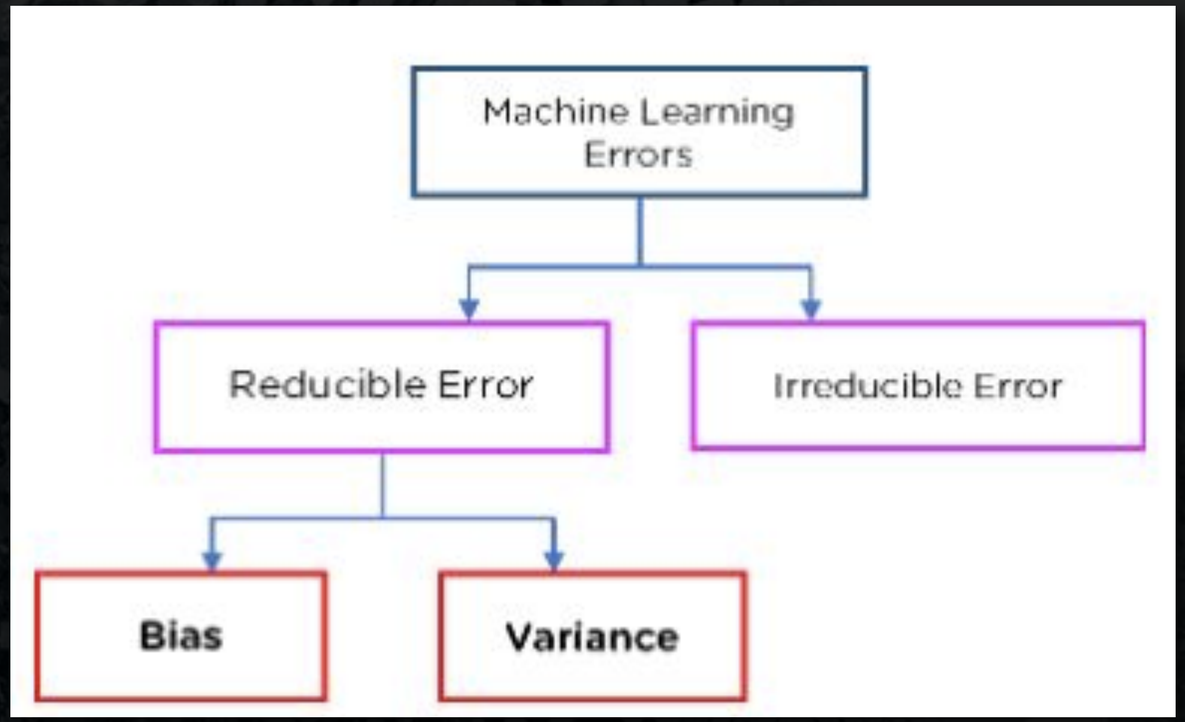
Over-fitting

Zero error on training data. The model is too sensitive only good for training dataset.
High deviation between training & test data

Prediction Error

The prediction error for any machine learning algorithm can be broken down into three parts:

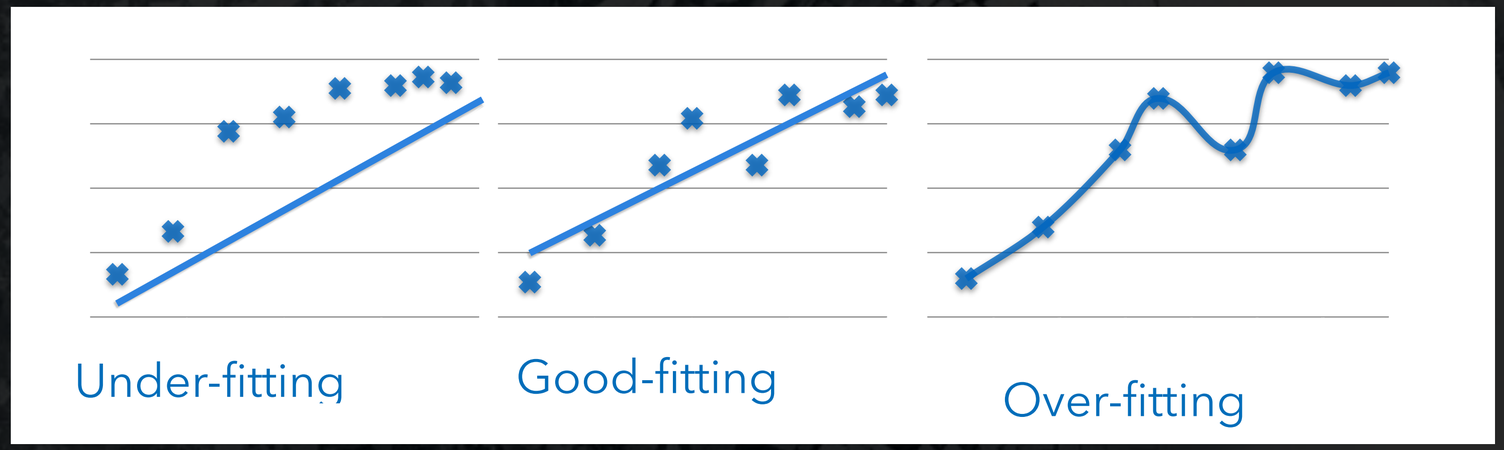
- Bias Error
- Variance Error
- Irreducible Error



The irreducible error cannot be reduced regardless of what algorithm is used. It is the error introduced from the chosen framing of the problem and may be caused by factors like unknown variables that influence the mapping of the input variables to the output variable.

Xilytica

Bias



Bias

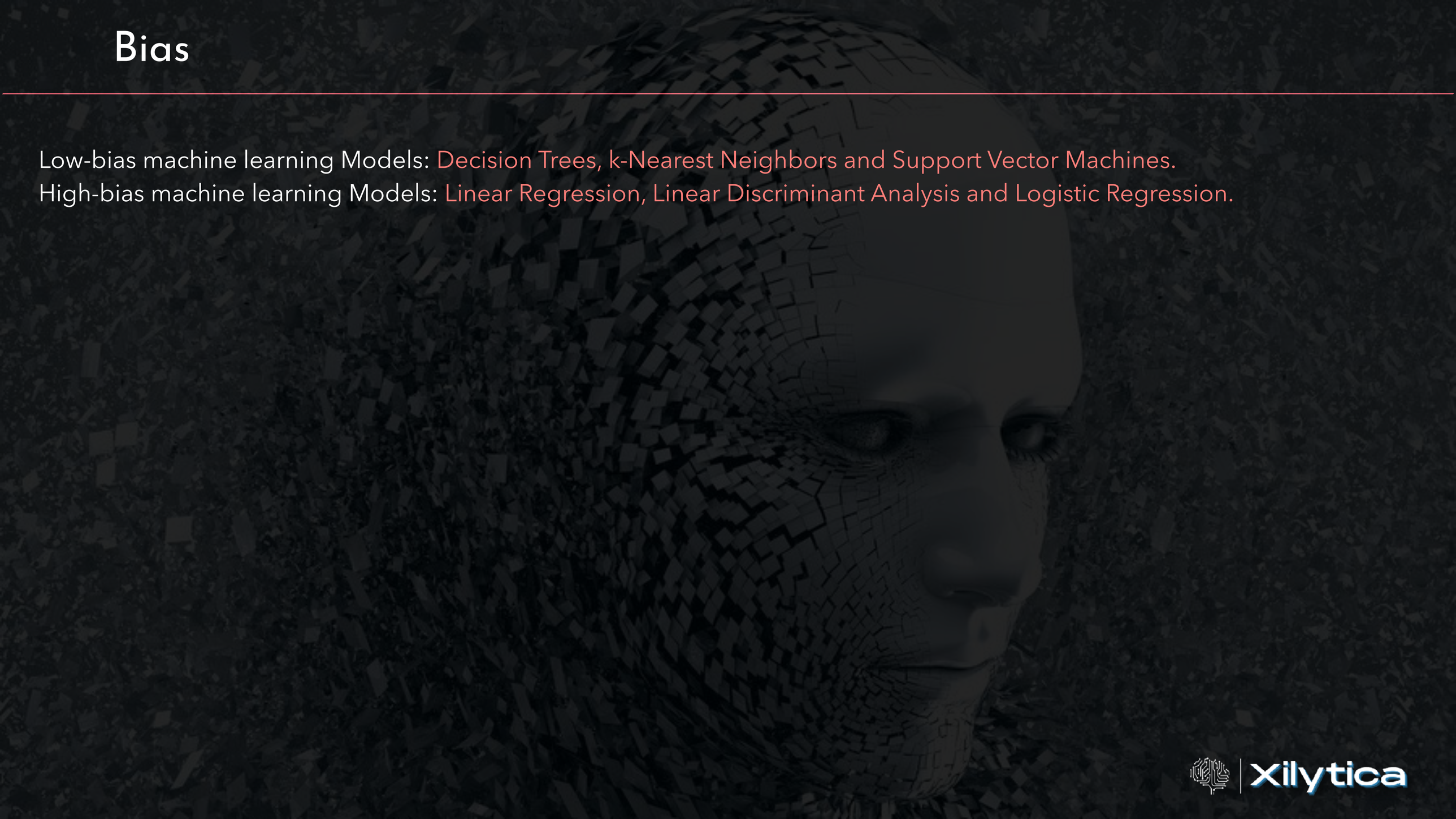
To make predictions, model needs to analyse the data & find patterns in the data, once the model is trained it is applied on the test or unseen data for prediction.

The bias is known as the difference between the predicted value and the actual value. When the Bias is high, assumptions made by our model are too basic, the model can't capture the important features of our data. This means that our model hasn't captured patterns in the training data and hence cannot perform well on the testing data too & can't perform better using test data as well. This instance, where the model cannot find patterns in our training set and hence fails for both seen and unseen data, is called Underfitting.

High Bias means a large error in training as well as testing data. Its recommended that an algorithm should always be low biased to avoid the problem of underfitting.

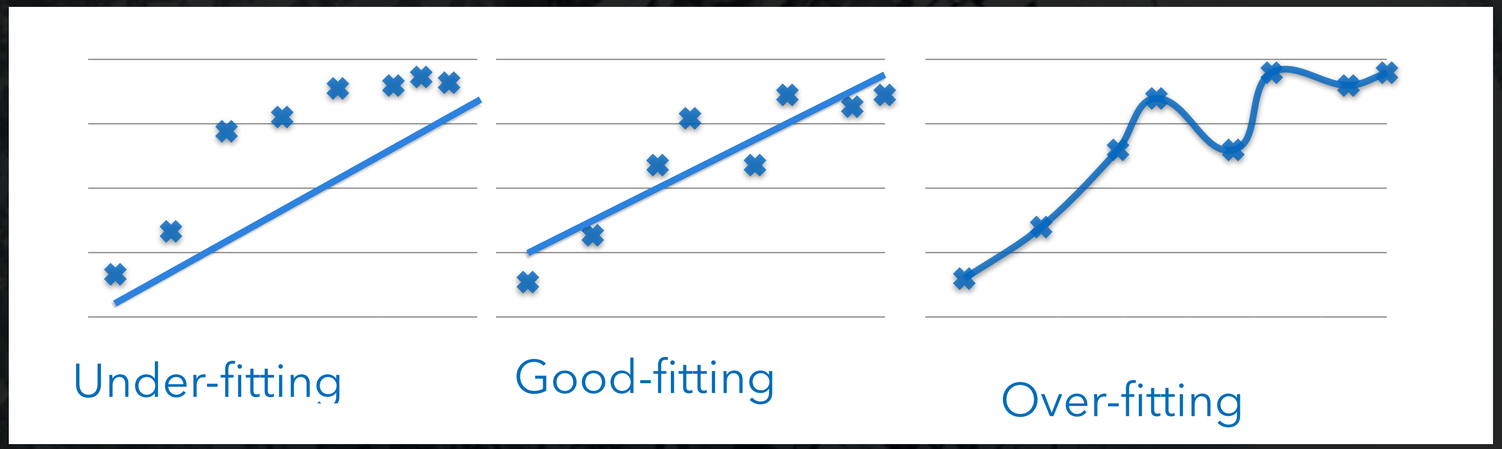
Bias

Low-bias machine learning Models: Decision Trees, k-Nearest Neighbors and Support Vector Machines.
High-bias machine learning Models: Linear Regression, Linear Discriminant Analysis and Logistic Regression.



Xilytica

Variance



Variance

Variance indicates how much the estimate of the target function will alter if different training data were used. In simpler terms the variability or inconsistency in the model prediction using different training dataset - it's not a measure of overall accuracy

In Machine Learning Model, high variance are strongly influenced by the specifics of the training data. This means that the specifics of the training have influences the number and types of parameters used to characterize the mapping function. Variance can lead to overfitting, in which small fluctuations in the training set are magnified.

Xilytica

Variance

Low Variance: Small changes to the estimate of the target function with changes to the training dataset.
High Variance: Large changes to the estimate of the target function with changes to the training dataset.

Generally, nonlinear machine learning algorithms that have a lot of flexibility have a high variance. For example, decision trees have a high variance, that is even higher if the trees are not pruned before use.

Low-variance machine learning Models: Linear Regression, Linear Discriminant Analysis and Logistic Regression.
High-Variance machine learning Models: Decision Trees, k-Nearest Neighbors and Support Vector Machines.

Bias & Variance TradeOff

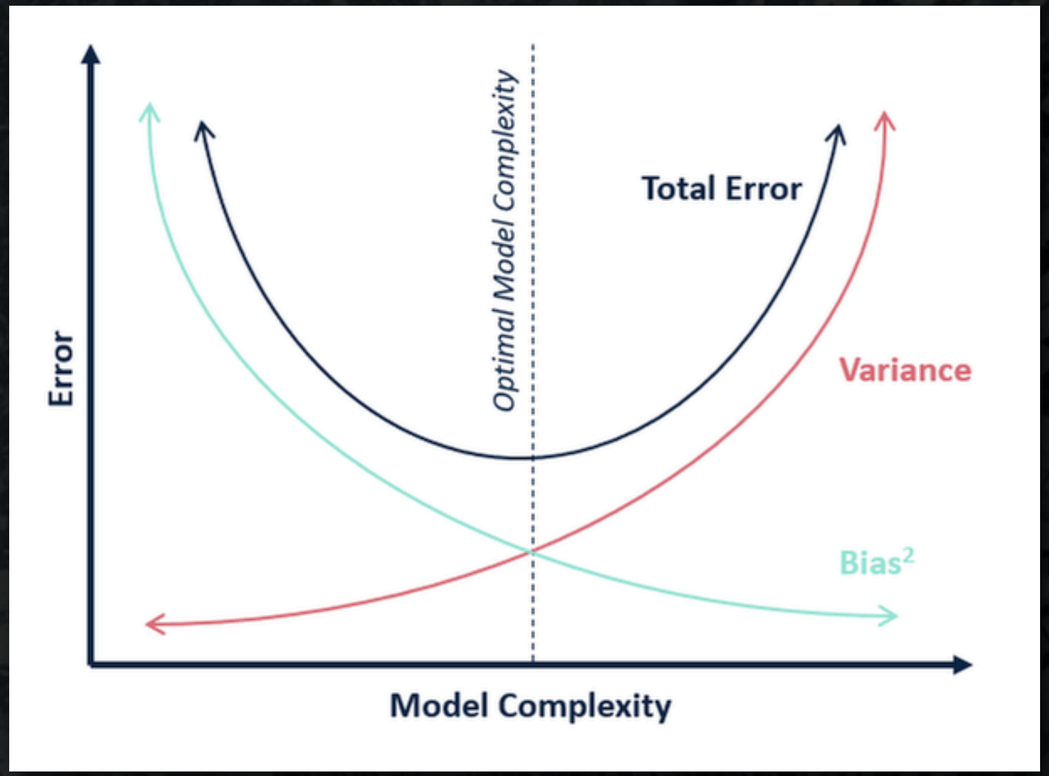
Objective of any supervised machine learning model is to achieve low bias & low variance in order to be a better predictive model or have a correct prediction power.

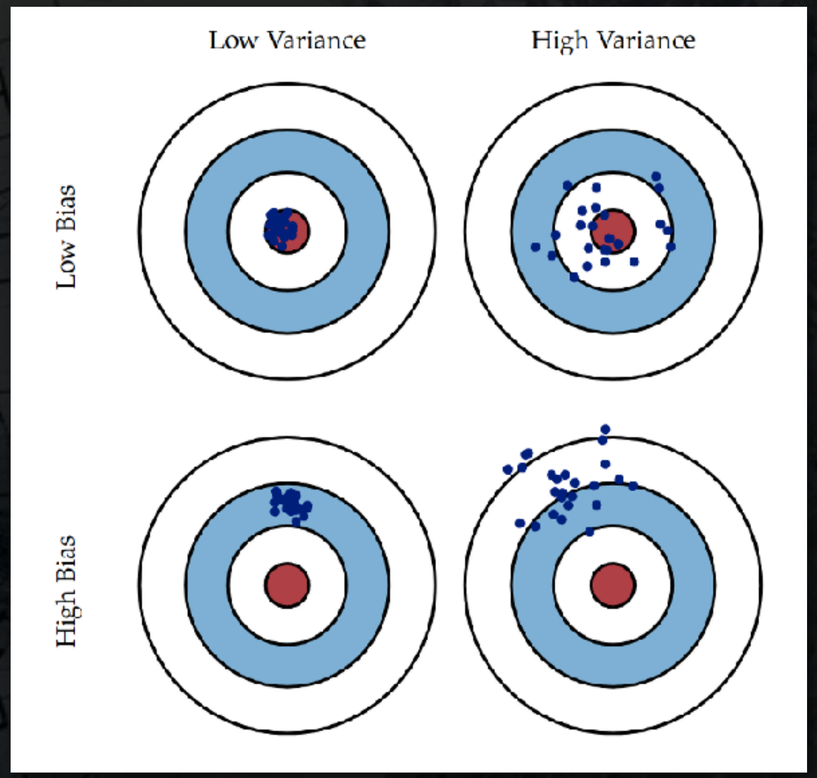
- Linear machine learning algorithms often have a high bias but a low variance.
- Nonlinear machine learning algorithms often have a low bias but a high variance.

Finding the right balance between the bias and variance of the model is called the Bias-Variance trade-off.

- Increasing the bias will decrease the variance
- Increasing the variance will decrease the bias

Bias & Variance TradeOff





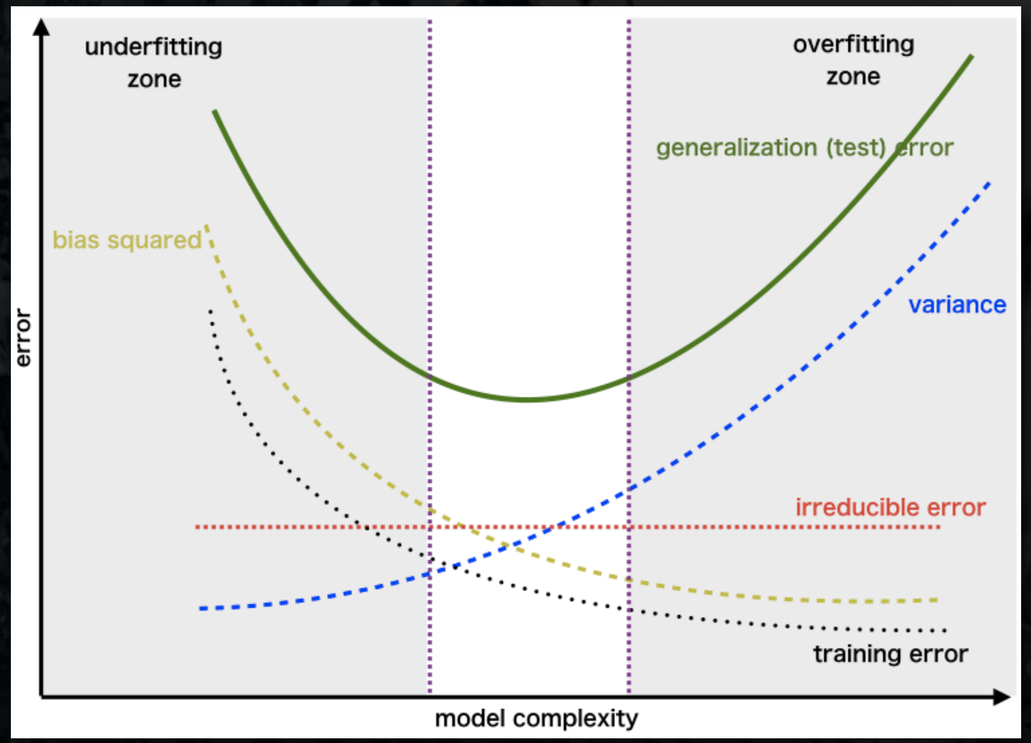
Finding the right balance between the bias and variance of the model is called the Bias-Variance trade-off.

- Increasing the bias will decrease the variance
- Increasing the variance will decrease the bias

Xilytica

Bias & Variance TradeOff

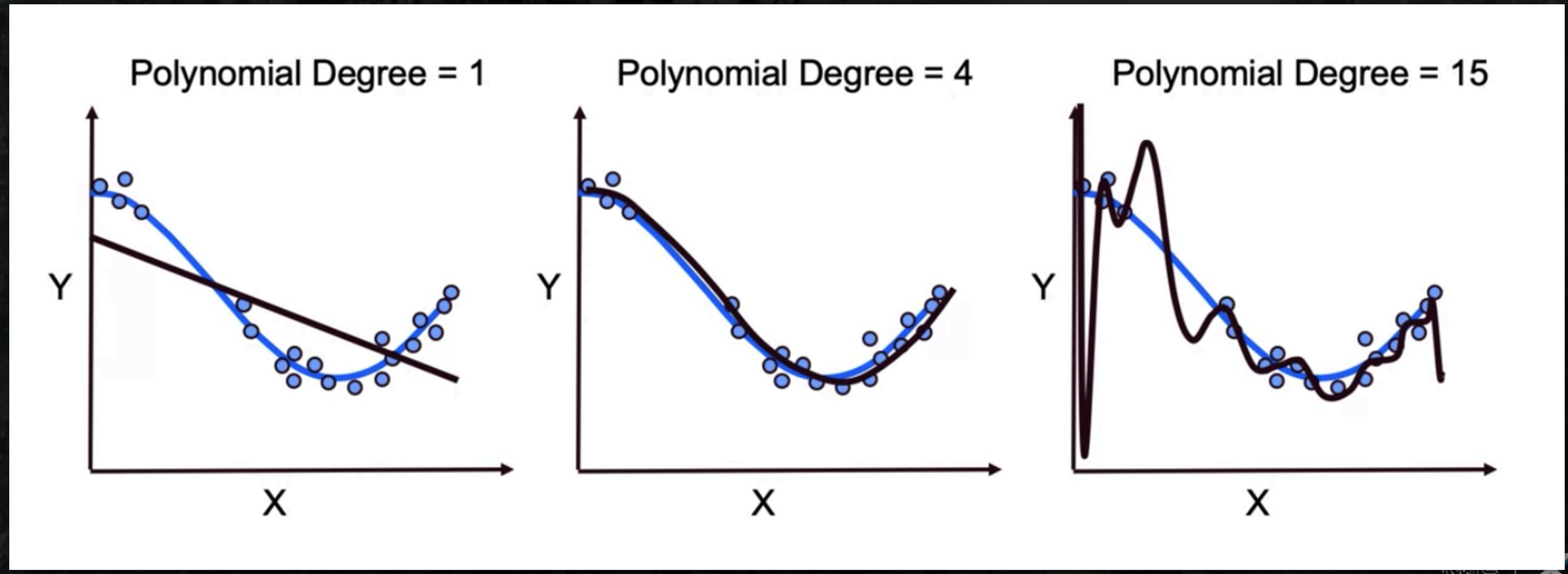
Finding the right balance with Low Bias and Low Variance is an iterative process where model is trained with different combination of features, Hyperparameters, different set of data set for training and test to find the right combination. We stop when we reach the point where Low Bias and Low variance is achieved and model is neither underfit and nor overfit.i.e Prediction accuracy is same on train and test data.



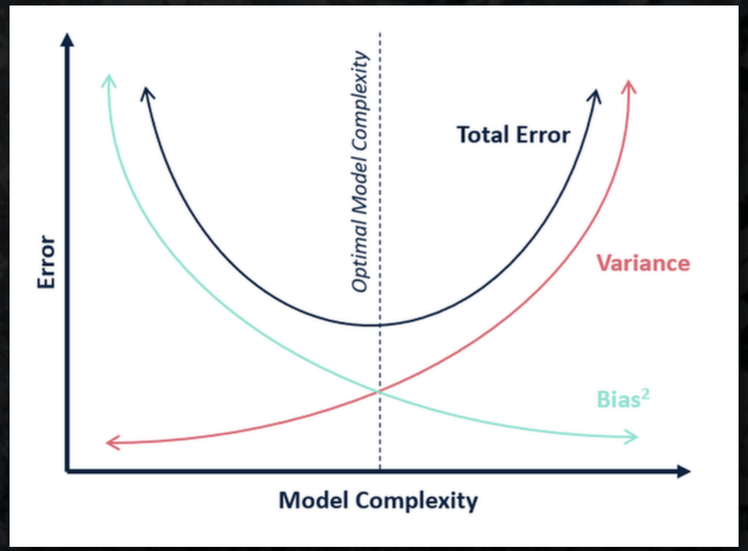
Xilytica

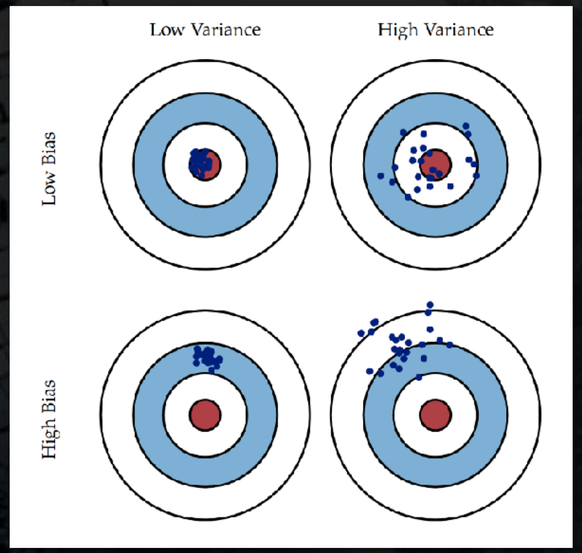
Bias & Variance TradeOff

- Higher the degree of polynomial regression, the more complex the model - Low Bias & High Variance
- At the lower degree, bias increases - predictions are too ridge to capture the curve pattern in the data
- At the higher degree, variance increases - predictions fluctuates widely because model's sensitivity
- The goal is to find the right degree, such that the model has the sufficient complexity to describe the data without overfitting



Bias & Variance TradeOff





Our goal is to minimize the total loss, which consists of bias, variance, and small noise. These curves show that increasing the complexity of the model, will decrease the bias, but the variance will increase and as a result, the total loss will be high. We can't take a too simple model, which can't even approximate the target function and can't take too big one either, because it has high variance. In the first case, our model will predict wrong results on the training set, cause the model is too simple, and in the second case our model will predict perfectly right results on the training set but will suffer on the examples from the validation set, on the examples, which the model hasn't seen in the training process. Here is how the predictions will look like

Dealing With High Bias & Variance

Lowering high Bias or Underfitting:

1. Use non Parameterised Algorithms
2. Make model more complex with more features
3. Use Non Linear Algorithms Example(Polynomial Regression, Kernel Function in SVM

Lowering high Variance or Overfitting:

1. Use More Data for training to make model learn maximum hidden pattern from the training data and model becomes generalised.
2. Use Regularization Techniques Example: L1 , L2, Drop Out, Early Stopping(in case of Neural Networks)etc.
3. Hyper Parameter Tuning to avoid Overfitting Example: Higher value of K in KNN, Tuning of C and Gama for SVM, Depth of Tree in Decision Tree
4. Use less number of features – Manual or Feature Selection Algorithms or automated using L1, L2 Regularization
5. Reduce complexity of Model – Reduce polynomial degree in case of Polynomial regression and Logistic regression
6. Use Advance techniques like Cross Validation, Stratified Cross Validation etc.

Regularisation - L1 & L2

One of the method to control or reduce complexity ignorer to prevent model to be overfit, is Regularisation.

Regularization - It essentially keeps all features, but reduces (or penalizes) the effect of some features on the model's predicted values. The reduced effect comes from shrinking the magnitude, and therefore the effect, of some of the model's term's coefficients.

$M(w) + \lambda R(w)$ - Adjusted Cost Function

$M(w)$ - Model Error - Mean Squared Error (MSE)

$R(w)$ - Functions of estimated parameter - Strengths or weight of different parameters

λ - Regularisation strength parameter, allows us to manage complexity tradeoff

$\lambda R(w)$ - penalise the model if its too complex

If stronger are the weights, the stronger our parameters, the higher the cost function is going to be & are goal is to minimise this by adding some bias to the model so we are not able to fit it closely to the actual training data. It is going to add the penalty, proportional to the size of the strength or weights of these parameters or functions of these parameters. larger the lambda is the more we penalise stronger parameters, which in turn makes our model to be less complex. This way we will increase the cost function based on how much do we want to penalise are model for being more complex.

Regularisation - L1 & L2

A higher lambda, introduces simpler model or more bias. So increasing lambda means more penalty for stronger weights, the more penalty is attributed to our weights, the less complex our model can be.

A lower lambda, means less regularisation that makes the model more complex & increase the variance

The two most popular regularization methods are ridge regression and lasso. Both methods involve adding a tuning parameter to the model, which is designed to impose a penalty on each term's coefficient based on its size, or effect on the model.

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

Cost function for ridge regression

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

Cost function for Lasso regression

Xilytica

Regularisation & Feature Selection

- Regularisation features as a feature selection by shrinking the contribution of features in a model. If we want to find out which all features are more important to the model we can use regularisation since it is going to take the contribution of each one of those features & eliminate or reduce them as it adds more weight to the penalty.
- With L1 regularisation, that is lasso regression, it can actually drive some of the coefficients or weights down to zero. Which means essentially removing the contribution of that feature altogether. Some times not all features are relevant for example in customer churn, the customer name probably does not add lot of value to the model. Also by eliminating few features model can be trained in lesser time.
- Feature elimination can also be used to identify what are the most important features, which can improve the model interpretability
- The two most popular regularization methods are ridge regression and lasso. Both methods involve adding a tuning parameter to the model, which is designed to impose a penalty on each term's coefficient based on its size, or effect on the model.

Regularisation - L1 & L2

One of the method to control or reduce complexity ignorer to prevent model to be overfit, is Regularisation.

Regularization - It essentially keeps all features, but reduces (or penalizes) the effect of some features on the model's predicted values. The reduced effect comes from shrinking the magnitude, and therefore the effect, of some of the model's term's coefficients.

The two most popular regularization methods are ridge regression and lasso. Both methods involve adding a tuning parameter to the model, which is designed to impose a penalty on each term's coefficient based on its size, or effect on the model.

Regularisations - Ridge Regression

The two most popular regularization methods are ridge regression and lasso. Both methods involve adding a tuning parameter to the model, which is designed to impose a penalty on each term's coefficient based on its size, or effect on the model. The larger the term's coefficient size, the larger the penalty, which basically means the more the tuning parameter forces the coefficient to be closer to zero. Choosing the value to use for the tuning parameter is critical and can be done using a technique such as cross-validation.

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

Cost function for ridge regression

Xilytica

Regularisations - Lasso Regression

The two most popular regularization methods are ridge regression and lasso. Both methods involve adding a tuning parameter to the model, which is designed to impose a penalty on each term's coefficient based on its size, or effect on the model. The larger the term's coefficient size, the larger the penalty, which basically means the more the tuning parameter forces the coefficient to be closer to zero. Choosing the value to use for the tuning parameter is critical and can be done using a technique such as cross-validation.

The Lass Regression can also work as a feature selection, this is due to the fact that the penalty term for each predictor is calculated slightly differently, and can result in certain terms becoming zero since their coefficients can become zero. This essentially removes those terms from the model, and is therefore a form of automatic feature selection.

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Cost function for Lasso regression

Xilytica

Assumptions of Linear Regression

- Linearity - y and x should be linear. Also y and parameters should be linear
- Homoscedasticity - Constant Variance in residuals
- Multivariate Normality - Check next slide
- Independence of Errors - IID normal (Independent & identically distributed)
    - Zero mean ,
    - Constant Variance - Homoscedasticity
    - Errors should be independent of each other if it not than we have auto correlation issue
- Lack of Multicollinearity - Variables should be Linearly independent of each other

Assumptions of Linear Regression

Linearity - y and x should be linear. Also y and parameters should be linear. Plot a scatter plot between y dependent variable & independent variables x1,x2,…xn. Scatter plot should look like a straight line. In case any of the variable on scatter plot shows a curve then using linear regression will not be a good fit.

Fixes -

Run a non linear regression or use a transformation methods like log transformation, exponential transformation etc

Assumptions of Linear Regression

No Endogeneity - Relationship between the independent variables & errors (difference between predicted values & observed values). Independent variables & errors are correlated. This is called omitted variables bias.

Omitted variable bias occurs when we forget to include a relevant variable in the model.

X be the independent variable & X* be the variable that we forgot to add in the model.
Both the variables are correlated with dependent variable Y.
That means the omitted variables is correlated with at least on independent X.

Anything that we can't explain in the model goes into the error. so error correlated with everything else. Omitted variable bias occurs when you forget to include a variable. This is reflected in the error term as the factor you forgot about is included in the error. In this way, the error is not random but includes a systematic part (the omitted variable).

Always look for correlation between independent variables & errors.

https://towardsdatascience.com/endogeneity-the-reason-why-we-should-know-about-data-part-i-80ec33df66ae

Assumptions of Linear Regression

Normality & Homoscedasticity of the Errors-

Normality - We assume that error term is normally distributed. For large samples central limit theorem applies to errors also. So we assume that normality exists when we have large dataset

Zero Mean of Error- If the mean is not zero it means the regression line is not the best fit. Having an intercept solves that problem. In real life it is unusual to violet this part of the assumption.

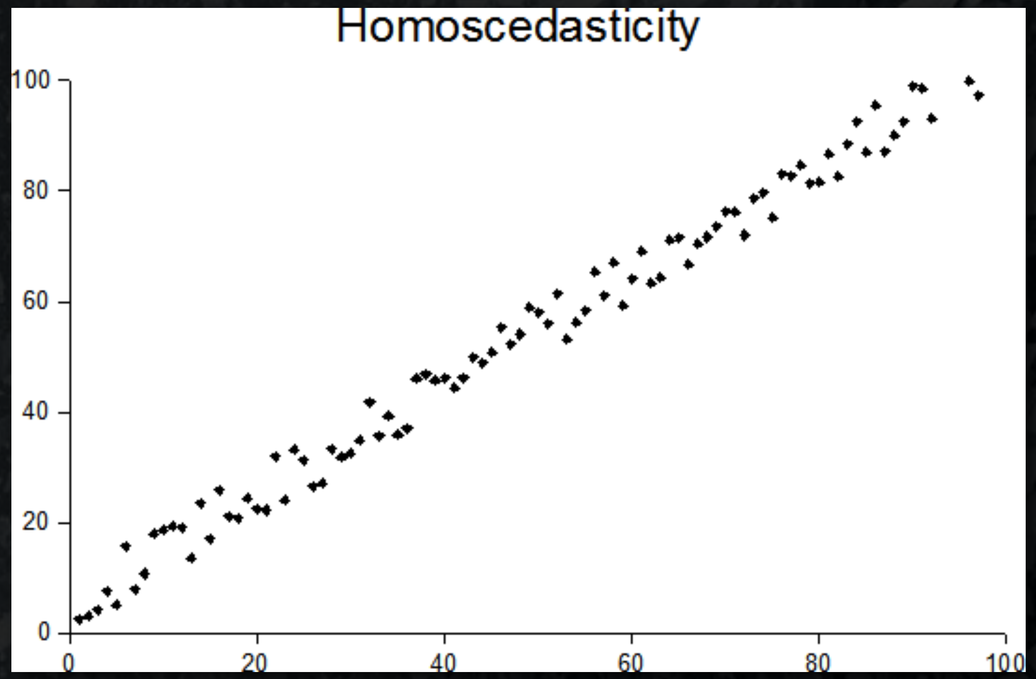
Homoscedasticity - Errors should have equal variance with other errors. If errors has a different variance and have pattern in the variance then it's a problem.

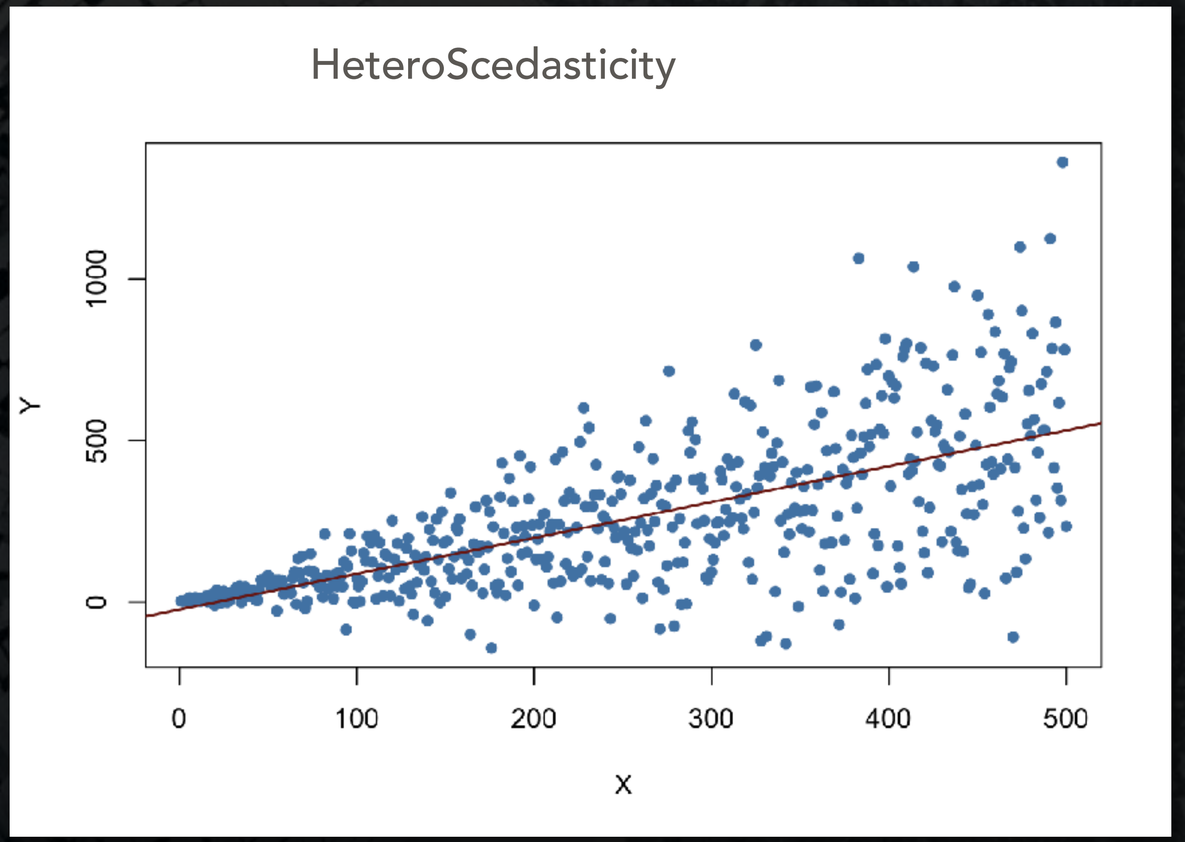
Fix -

- *Look for Omitted Variable Bias*
- *Look for Outliers*
- *Try to transform data - Log or exponential or any other*

    - y hat = b0 + b1log X1 (Semi log model)
    - log y hat= b0 + b1logX1 (Log - Log model) - That means As X increase by 1 percent Y increases by b1 percent
    - Log y hat = b0 + b1X1 (Semi Log model) - That means - As X increase by one unit Y increases by b1 percent

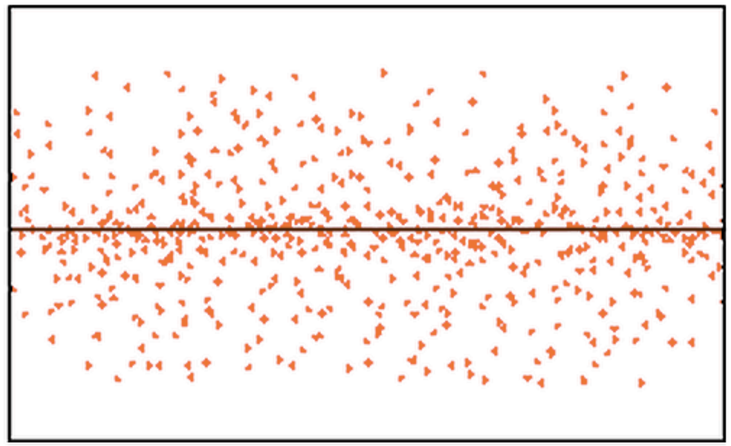
Assumptions of Linear Regression



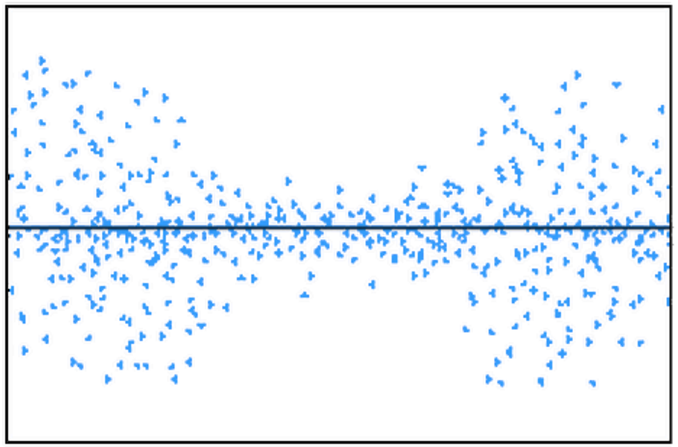


Xilytica

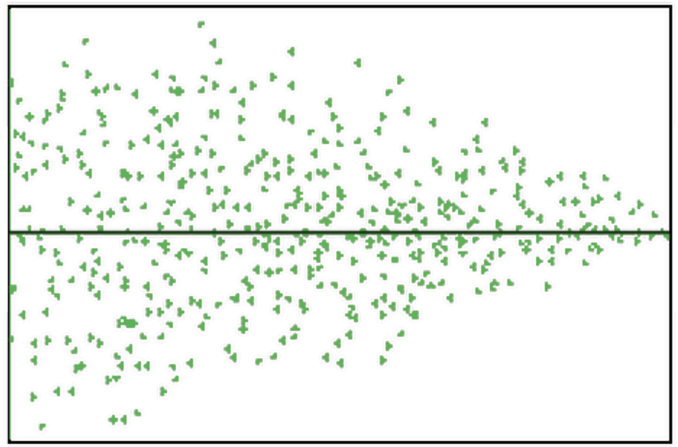
Assumptions of Linear Regression



Random Cloud (No Discernible Pattern)



Bow Tie Shape (Pattern)



Fan Shape (Pattern)

Assumptions of Linear Regression

No Autocorrelation - or No serial correlation

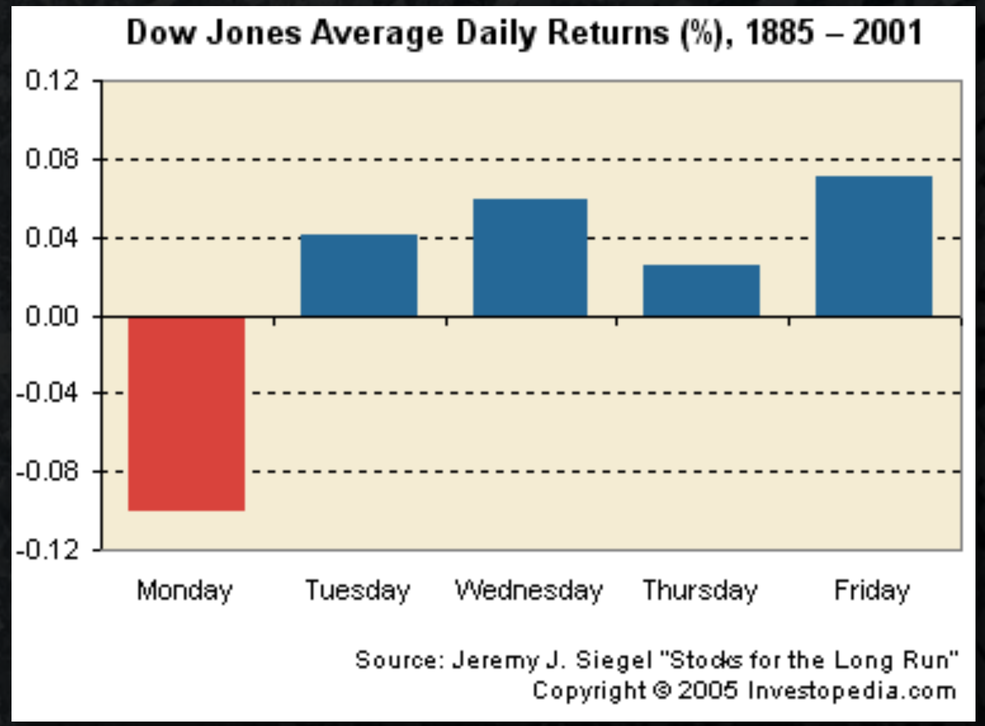
- Errors are assumed to be uncorrelated . Not very easy to find in cross sectional data but very easy to find in time series data. Eg. stock prices. In time series for same stock ups and down should be dependent on GDP, tax rate, political events, etc. but there is pattern in stock prices. Day of the week effect which means high returns on Fridays and low returns on Mondays.

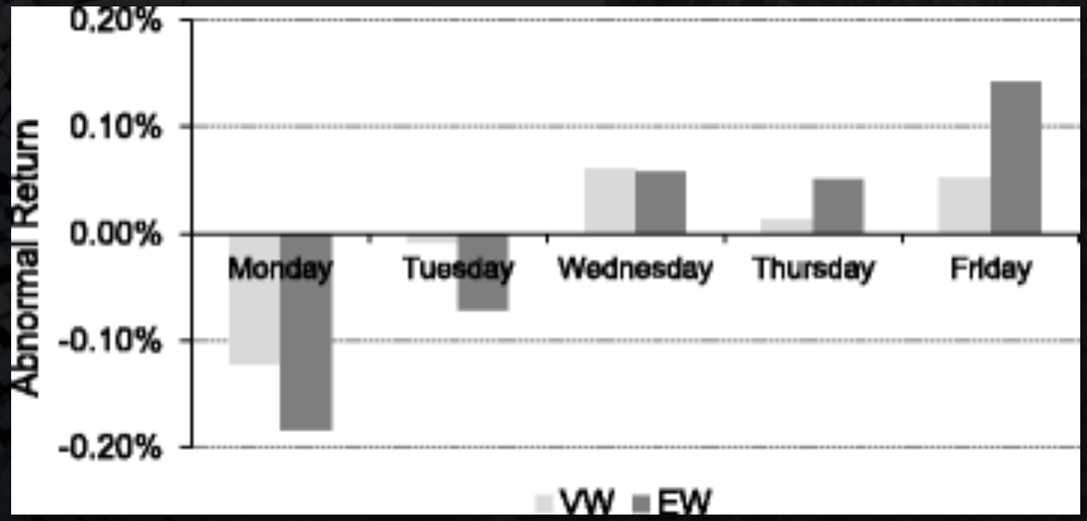
- Investors read news over the weekend and based on that they sell stocks on Mondays and when they get new positive information during the week they buys on Thursdays & Fridays. Errors on Mondays will be biased downwards and error on Fridays will be biased upwards

- Check for pattern in the errors. If their is no pattern in the errors then their is no auto correlation. Plot residuals and try to find pattern.
- Durbin Watson Test - Values fall generally between 0 and 4. If the value is 2 it means no auto correlation. If value is less than 1 or more than 3 then its a cause for an alarm.

- There is no remedy as such but to use any other regression model for time series data such as Auto regressive model, moving average model, autoregressive moving average model or Autoregressive integrated moving average model.

Assumptions of Linear Regression





Xilytica

Assumptions of Linear Regression

No Multi-collinearity

We see high multi collinearity we we have two or more variables have high correlation.

a = 3 + 4*b or
b = (a - 3)/4

Here a can be represented using b. Hence models containing a & b we have multicollinearity of 1. This causes a big problem to our regression model and coefficients will be wrongly estimated. If B can be represented using A then their is no point in using both.
If we have multicollinearity of 0.9 then this is also a problem with the model.

Fix -
Drop one of the two variables
Transform them into one (eg. by taking average)
Keep them both but be very cautious

Prevention -
Find correlation between each of the two pairs of independent variables.