# **Regression Analysis**

Factor Analysis 🡪 Academic Papers 🡪 Machine Learning

Amount of Money you spend 🡪 Amount of money you earned

1. Simple Linear Regression
2. Multiple Linear Regression
3. How to build a Regression
4. How to interpret it
5. How to Compare Different Models

|  |  |  |
| --- | --- | --- |
| **Library** | **Used for** | **Used in** |
| import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  import seaborn as sns  sns.set() | Common | Common |
| from sklearn.cluster import KMeans | Cluster Analysis | Cluster Analysis |
| from sklearn.linear\_model import LinearRegression | Linear Regression | Linear Regression |
| from sklearn.preprocessing import StandardScaler | Standardization | Common |
| from sklearn.model\_selection import train\_test\_split | Divide the data into train and test | Common |
| import statsmodels.api as sm | Summarize the Stats | Linear & Logistic |
| from statsmodels.stats.outliers\_influence import variance\_inflation\_factor | To calculate the Variance Influence Factor | Linear & Logistic |
| from sklearn.feature\_selection import f\_regression  f\_regression(x,y)  p\_values = f\_regression(x,y)[1] | To find out p-values and it’s respective F- Values | Linear & Logistic |

# **Linear Regression**

A linear regression is a linear approximation of a casual relationship of casual relationship between two or more variables.

Process:-

Get Sample Data 🡪 Design a model that works for that sample 🡪 Make predictions for the whole populations

Dependent Variable (Y) and Independent (Predictors – x1, x2, …………… xn)

Simple Linear Regression Model -> y = β0 + β1x1 + ε

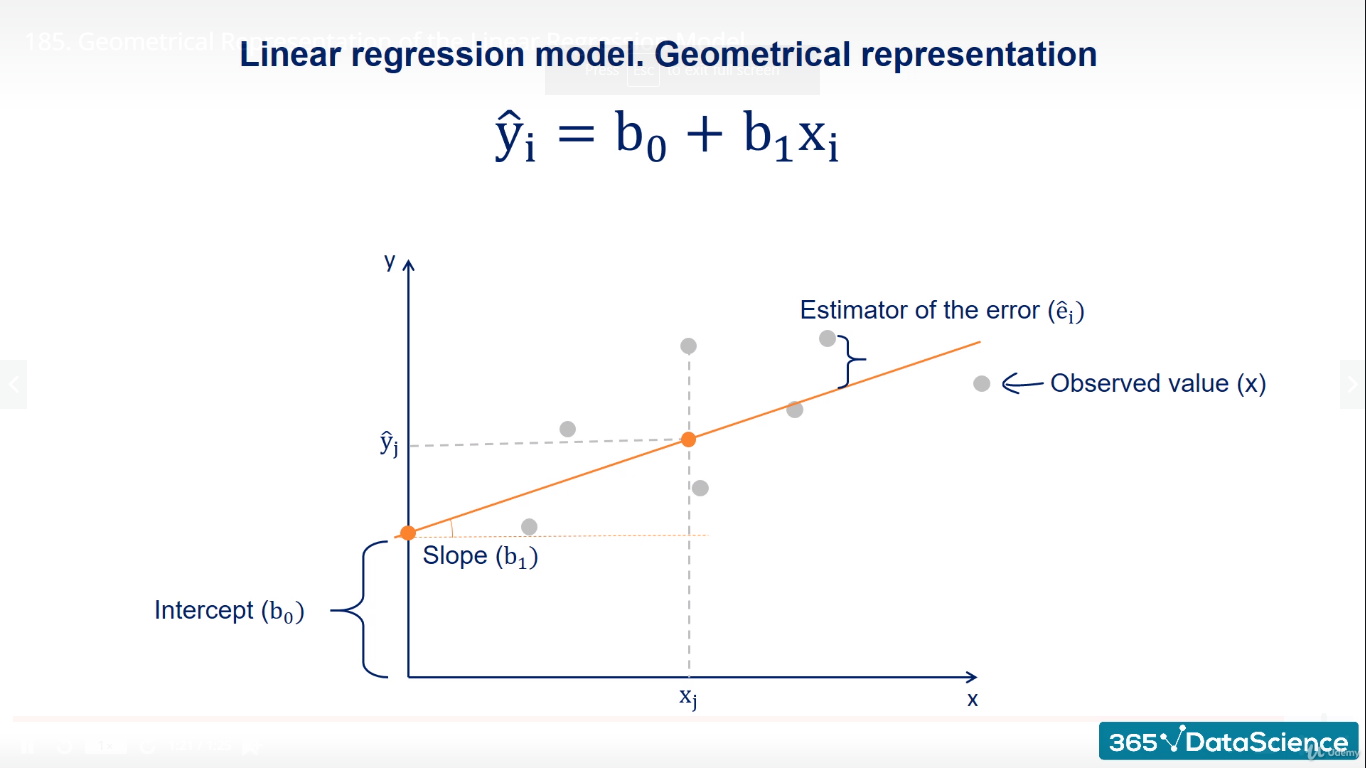
y- Dependent Variable, x1 – independent variable, β0 – constant, ε – Error of estimation

ŷ = β0 + β1x1

ŷ – Estimated/predicted value, b0 – constant, b1 – Quantified the effect the independent (x) on the dependent (y), x1 – sample data for independent variable

**Correlation vs Regression**

|  |  |
| --- | --- |
| **Correlation** | **Regression** |
| Relationship | One variable affects the other |
| Movement Together | Cause and effect |
| p(x,y) = p(y,x) | One way |
| Single Point | Line |



**Packages Used:**

Open Anacond Prompt and enter conda list

Conda install seaborn

Pip install seaborn

Import numpy as np

Import pandas as pd

Import matplotlib.pyplot as plt

Import scipy

Import seaborn as sns

Import statsmodels.api as sm

Import sklearn

1. **Numpy :-** a third party package allowing us to work with multidimentional arrays.
2. **Pandas:**- allows us to organize data in a tabular form and to attach descriptive labels to the rows and columns.
3. **Matplotlib:** a 2D plotting library specially designed for visualization of NumPy computations.
4. **SciPy:** is a python ecosystem containing of lot of tool for scientific calculation suitable for Mathematics, Machine Learning and Engineering. Scify is a group of Numpy, pandas and Matplotlib library.
5. **Seaborn:** Seaborn is a Python Visualization library based on matplotlib. It provide a high level interface for drawing attractive statistical graphics. Seaborn is the package build top on the matplotlib.
6. **Statsmodels:** A package with on top of Numpy and Scipy, which integrates with pandas. SM provides very good summaries.
7. **Scikit-learn**: sklearn is one of the most widely used machine learning libraries.



A screenshot of a cell phone

Description automatically generated

**Standard Error:-** shows the accuracy of the prediction. The lower the standard error the better is the accuracy.

**t-statistic and it’s P value (P>(t)):-** p value is less than 0.05 then the variable is significant.

**Variability:**-

A close up of a map

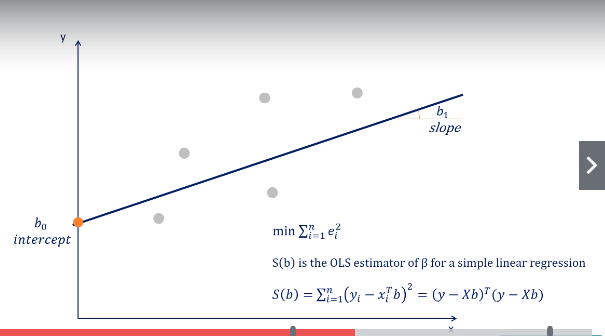
Description automatically generated

Dependent Variable – GPA

Model – OLS (Ordinary Least Squares)

**OLS (Ordinary Least Squares)** :-

Lower error 🡪 better explanatory power 🡪 Best explanatory power



**Other Methods**

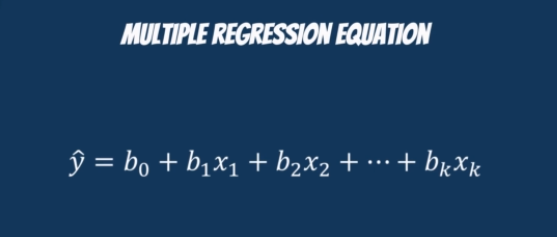
1. Generalized Least Squares
2. Maximum likelihood estimation
3. Bayesian regression
4. Kernel Regression
5. Gaussian Process Regression

**R-Squared:- measure the goodness of fit.**

Total Variability (SST) fall into two categories a. Explained (SSR) b. Unexplained (SSE)

Is equal to Variability explained by the regression (SSR) divided by Total variability of the dataset (SST)

# **Multiple Linear Regression**



It’s about the best fitting model. Min SSE

**R Squared:**

1. The R Squared is measured how much of the total variability is explained by our model.
2. Multiple regressions are always better than simple ones, as with each additional variable you add, the explanatory power may only increase or stay the same

**Adjusted R Squared:** measures how well your models fits the data. However its’s penalizes the use of variables that are meaningless for the regression.

If we add a new variable with random numbers. It will not impact our final results.

A screenshot of a cell phone

Description automatically generated

**F – Statistic:** is used for testing the overall significance of the model. The lower the F-Statistic, the closer to a non-significant model

**OLS – Assumptions:**

A picture containing screenshot

Description automatically generated

**Linearity:** The Linear regression is the simplest non-trivial relationship

**Fixes:**-

* Run a non-linear regression
* Exponential transformation
* Log transformation

**Endogeneity of Regressors:** Errors means predicted values to actual values correlated with independent variable

A screenshot of a cell phone

Description automatically generated

Omitted Variable Bias

1. Where did we draw the sample from? **Central London**
2. Can we get a better sample? **Big enough**
3. Why is bigger real estate cheaper? **Not in the city of London**
4. What is it about smaller size that is making it so expensive? **Location**
5. Where are the small houses? **City of London**

Omitted Variable bias occurs when you forget to include a variables. This is reflected in the error term as the factor you forget about is included in the error.

**Normality and Homoscedasticity:-** ε α N(0, σ2)

* Normality (Error term normally distributed). T-tests and F-tests work because we have assumed normality. For Large dataset Central Limit Theorem applied for error terms too.
* Zero Mean :- having an intercept solves that problem
* Homoscedasticity :- equal variance one with the others

Prevention:-

* Look for Omitted Variable Bias
* Look for outliers
* Transform :- Take the log of the variable, perform the regression.

**No Autocorrelation:- (no serial correlation)**

**Durbin-Watson Test:-** 2 🡪 No autocorrelation, <1 and >3 cause an alarm, No Remedy

Alternatives:-

* Autoregressive Model
* Moving average model
* Autoregressive moving average model
* Autoregressive integrated moving average model

**No Multicollinearity:** two or more variable has high correlation.

Sklearn does not have this functionality so we will use statsmodel. One of the best ways to check for multicollinearity is through VIF (Variance Inflation Factor)

Multicollinearity is a common problem when estimating linear or generalized linear models, including logistic regression and Cox regression. It occurs when there are high correlations among predictor variables, leading to unreliable and unstable estimates of regression coefficients. Most data analysts know that multicollinearity is not a good thing.  But many do not realize that there are several situations in which multicollinearity can be safely ignored.

**VIF (Variance Inflation Factor):-** Diagnostic for multicollinearity, the variance inflation factor (VIF). The VIF may be calculated for each predictor by doing a linear regression of that predictor on all the other predictors, and then obtaining the *R2* from that regression. The VIF is just 1/(1-*R2*).

|  |
| --- |
| from statsmodels.stats.outliers\_influence import variance\_inflation\_factor  variables = data\_cleaned[['Mileage', 'EngineV']]  vif = pd.DataFrame()  vif["VIF"] = [variance\_inflation\_factor(variables.values, i) for i in range(variables.shape[1])]  vif["features"] = variables.columns |

If vif = 1 🡪 No Multicollinearity, 1 to 5 🡪 perfectly okay, above then 6 unacceptable

Ƿab = 1 (perfect multicollinearity), Ƿcd = 0.9 (imperfect multicollinearity)

Rationale: if c can be almost represented using d, there’s no point in using both

Fixes:

* Drop one of the two variables
* Transform them into one (e.g. average price)
* Keep them both (if you are super confidence)

Preventation:-

* Find the correlation between each two pairs of independent variables Ƿxi, xj for I,j; i≠ j

**Dummy Variable:-** SAT, GPA along with Attendance with Categorical Data and we can’t put into the regression. We would map the ‘Yes’ or ‘No to 1 and 0.

Imitation of categories with numbers.



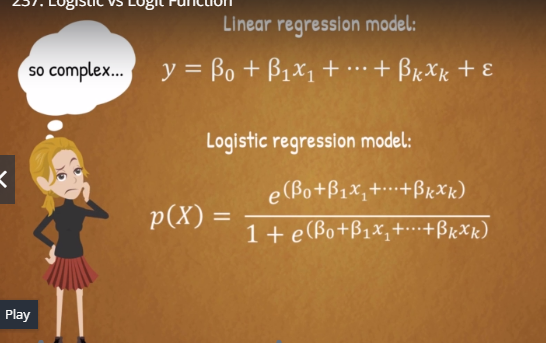
# **Logistic Regression**

The Logistic regression predicts the probability of an event occurring.

Input 🡪 Probability

A screenshot of a social media post

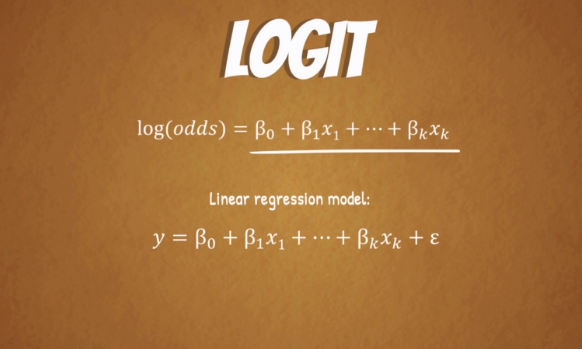
Description automatically generated



A picture containing text

Description automatically generated

**Logit Model**



Logistic Regression Assumptions

1. Linearity = Non Linear
2. No endogeneity
3. Normality and homoscedasticity
4. No autocorrelation
5. No Multicollinearity

|  |
| --- |
| Importing the Libraries  Import numpy as np  Import pandas as pd  Import matplotlib.pyplot as plt  Import statsmodels.api as sm  Import seaborn as sns  Sns.set() |
| Upload the Data  data = pd.read\_csv('2.01. Admittance.csv') |
| Data Mapping  data['Admitted'] = data['Admitted'].map({'Yes':1,'No':0}) |
| x1 = data['SAT']  y = data['Admitted'] |
| x = sm.add\_constant(x1)  x.head() |
| reg\_log = sm.Logit(y,x)  results\_log = reg\_log.fit()  Notes: Optimization terminated successfully. Current function value: 0.137766. Iterations 10 |
| results\_log.summary() |

SM uses a machine learning algorithm to fit the regression

Function value : value of the objective function at the 10th Iteration

A screenshot of a cell phone

Description automatically generated

**MLE (Maximum likelihood estimation)**

**Likelihood function:** a function which estimates how likely it is that the model at hand describe the real underlying relationship of the variables.

The bigger the likelihood function, the higher the probability that our model is correct!.

MLE tries to maximize the likelihood function.

The computer is going through different values, until it finds a model, for which the likelihood is the highest.

When it can no longer improve it. It will just stop the optimization.

**Log Likelihood: same as above with Log function.**

**LL Null (log likelihood- null):** the log-likelihood of a model which has no independent variables. (y = b0)

You may want to compare the log-likelihood of your model with the LL-null, to see if your model has any explanatory power

**LLR (log likelihood ratio):** measure if our model is statistically different from LL-null, a.k.a. a useless model

**Pseudo R-Squared:**

**McFadden’s R -Squared:** A good Pseudo R-Squared is somewhere between 0.2 and 0.4. This measure is mostly useful for comparing variations of the same model. Different models will have completely different and incomparable Pseudo R-Squares!

A screenshot of a cell phone

Description automatically generated

**Confusion Matrix:**

**A screenshot of a cell phone

Description automatically generated**

For 69 observations the model predicted 0 and the true value was 0

For 90 observations the model predicted 1 and the true value was 1

In 159 cases the model did its job well

For 4 observations the model predicted 0 while the true value was 1

For 5 observations the model predicted 1 and the true value was 0

In 9 cases the model ‘got confused’

Overall the model made an accurate prediction in 159 out of 168 cases. (94.6%)

# **Cluster Analysis**

Cluster analysis is multivariate statistical technique that groups observations on the basis some of their features or variables they are described by

Observations in a dataset can be divided into different groups and sometimes this is very useful

A screenshot of a cell phone

Description automatically generated

The goal of clustering is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters.

**Cluster Analysis:** Expore the data and identify patterns

1. Several clustering problems
2. How to perform cluster analysis
3. How to find the optimal number of clusters
4. How to identify appropriate features
5. How to interpret results

**Applications**

1. Market Segmentation
2. Image Segmentation

A picture containing text

Description automatically generated

**Supervised Learning**

**Labelled data**

**Inputs**

**Correct Values for outputs**

**Model (Inputs) 🡪 Outputs 🡪 Correct values for outputs**

**Logit(SAT,Gender) 🡪 Predictions 🡪 Admitted Data**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**A close up of a map

Description automatically generated**

**A screenshot of a social media post

Description automatically generated**

**Euclidean distance: W**hen performing clustering we will be finding the distance between clusters (n-dim Space).

**Centroid:** A centroid is the mean position of a group of points (ake center of mass).

**K-Mean Clustering:**

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids
5. Repeat 3 and 4

A screenshot of a cell phone

Description automatically generated

A seed is basically a starting centroid

**Selecting the Number of Clusters:**

**Elbow method:**

1. Minimizing the distance between points in a cluster
2. Maximizing the distance between clusters

**WCSS:**

Distance between points in a cluster – ‘within-cluster sum of squares’, or WCSS

Similar to SST, SSR and SSE. WCSS is a measure developed within the ANOVA framework (SST = SSR+SSE)

If we minimize WCSS, we have reached the perfect clustering solution.

A close up of a sign

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Pros and Cons of K-Means:**

|  |  |  |
| --- | --- | --- |
| **Pros** | **Cons** | **Remedies** |
| Simple to understand | We need to pick k | The Elbow method |
| Fast to cluster | Sensitive to initialization | k-means ++ |
| Widely available | Sensitive to outliers | Remove outliers |
| Easy to implement | Produces spherical solutions |  |
| Always yields a result | Standardization |  |

**To standardize or not to standardize?**

**A screenshot of a cell phone

Description automatically generated**

The ultimate aim of standardization is to reduce the weight of higher numbers.

Every rectangle on a graph (only 4 points) after being standardized turns into a square

**Relationship between Regression and Cluster:**

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Types of Analysis**

1. Exploratory
   1. Get acquainted with the data
   2. Search for patterns
   3. Plan
2. Confirmatory
3. Explanatory

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Types of Clustering**

1. **Flat (K – Means)**
2. **Hierarchical (Taxonomy of the Animal Kingdom)**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**A picture containing device

Description automatically generated**

**Dendogram:** is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

**A screenshot of a social media post

Description automatically generated**

**A close up of a logo

Description automatically generated**

**Heatmaps:** A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. “Heat Map” are represented as colors. ‘Heat Map” is a newer term but shading matrices have existed for over a century.