What is Standardization / Feature scaling?

What is evaluation metrics for regression?

What is Mean Absolute Error (MAE)

This is the mean of the absolute value of errors

What is Squared Error (MSE)

This is the mean of the squared errors, Larger errors are noted more than with MAE, making MSE more popular.

What is Root Mean Square Error (RMSE)

This is the root of the mean of the squared errors, most popular (has same units as y)

What is R Squared Values

What is ROC curve?

What is confusion matrix? - classification problem

What is overfitting?

What is underfitting?

What is Dummy Variables?

What is log transformation?

What is Multicollinearity?

What is centroid?

What is Euclidean distance?

What is Elbow method?

What are the types of Analysis?

Exploratory –

Get acquainted with the data

Search for patterns

Plan

Confirmatory -

Explanatory -

What is odds? of getting Head in tossing a coin

exp(ln(x)) = |x|

log(exp(x)) = x

Linear Reg:

p-value :

p-value <0.05, means that the variable is significant, therefore the co-efficient (intercept, slope) are not zero

:

measures how much of the total variability is explained by our model, value comes between 0-1, Higher the value better the variable explained in the model.

Adjusted

Adjusted penalizes excessive use of variables, so Adjusted <

when we add new variable in the existing model, the and Adjusted value should increase. if the value increased and Adjusted value decrease, then the variable can be omitted since it holds no predictive power

F- Statistic:

Lower the F-Static closer the non-significant model (Prob(F-statistic))

Logistic Reg:

MLE:

Maximum Likelihood Function, the bigger the likelihood function, the higher the probability that our model is correct

Log-Likelihood:

The value of the log likelihood is almost but not always negative, bigger the better

Simple Linear Regression Model (causal relationship)

y - Dependent variable

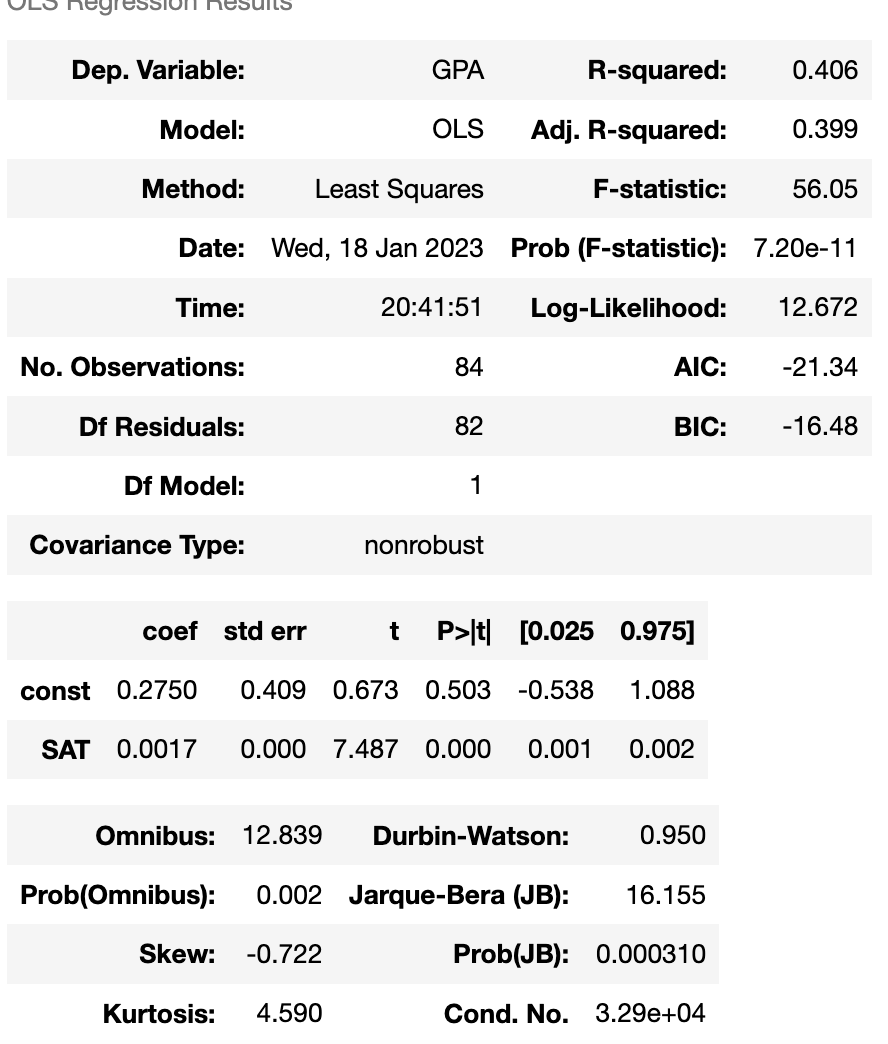
- Independent variable

- co-efficient (slope)

- Constant (intercept)

- Error

|  |  |
| --- | --- |
| CORRELATION | REGRESSION |
| Relationship | One Variable affects the other |
| Movement together | Cause and Effect |
| p(x,y) = p(y,x) | One Way |
| Single Point | Regression Line |



A p-value <0.05, means that the variable is significant, therefore the co-efficient (intercept, slope) are not zero

What does a p-value of 0.503 suggest about the intercept coefficient?

It is not significantly different from 0 (as such we are interested in in depended variable, we don’t take this (intercept coefficient - p value)

What does a p-value of 0.000 suggest about the coefficient of x?

It is significantly different from 0

Sum of Squares Total SST . Measures the total variability of the dataset (TSS)

Sum of Squares Regression SSR Difference between predicted value and the mean value, Measures the explained variability by your line (ESS)

Sum of Squares Error SSE Difference between the observed value and the predicted value, Measures the unexplained variability by the regression (RSS)

SST = SSR + SSE = +

**Different Kinds of Regression Model**

Generalized least squares

Maximum likelihood estimation

Bayesian Regression

Kernel Regression

Gaussian Process Regression

– Measures Goodness of Fit (higher the correct variable / factor better the )

=

= 0, Regression explains NONE of the variability

= 1, Regression explains the entire variability

= 0.5, Regression explains the half of the variability, should consider adding more variable to the regression

# measures how much of the total variability is explained by our model! .

# 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 penalizes excessive use of variables, so Adjusted <

# value comes between 0-1, Higher the value better the variable explained in the model.

# when we add new variable in the existing model, the and Adjusted value should increase. if the value increased and Adjusted value decrease, then the variable can be omitted since it holds no predictive power.

F- Statistic

# Lower the F-Static closer the non-significant model (Prob(F-statistic))

In Regression What is

1. Correlation
2. Causation
3. Simple LR Model
4. Multivariate LR Model
5. Geometrical Representation
6. SST, SSR, SSE
7. OLS
8. R-Squared
9. Adjusted R-Squared
10. F-Test

Assumptions:

1. Linearity
2. No Endogeneity = 0:
3. Normality and Homoscedasticity )
4. No Autocorrelation
5. No Multicollinearity

How to choose Linear Regression Model (:

# The relation between the dependent and independent variable goes through the linear line, then the model is good for LR, but if the variable dose not lined linear line then this is not good for Linear Regression. What should do if you want to employ a linear regression but the relationship in your data in not – Transform it appropriately before using it.

Fixes:

1. Run a non-linear regression
2. exponential transformation (exponential/polynomial – in to linear)
3. log transformation

No Endogeneity = 0:

– Error Is difference between the Absorbed value and the predicted value, if this is correlated with independent value then its referred to omitted variable bias

Omitted Variable Bias are - always different , always sneaky, only experience and advanced knowledge can help

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 included a systematic part

Normality and Homoscedasticity )

To prevent High Variation, we can do Look for OVB, Look for outliers, Transform

What is Elasticity? As X increases by 1 percent, Y increases by b1 percent - ( Log – Log Mode log

No Autocorrelation

\* If there is a Autocorrelation it’s better to avoid regression

Durbin-Watson from Stats Model - in general the value falls between 0 and 4, 2 indicates no Autocorrelation, < 1 and > 3 cause an alarm

Remedy may be

1. Autoregressive model
2. Moving average model
3. Autoregressive moving average model
4. Autoregressive integrated moving average model

No Multicollinearity Two or More variable has high correlation

Fixes: Drop one of the two variables, Transform them into to one ( eg. average price)

Prevention: Find the correlation between each two pairs of independent variables

What is Dummy Variables?

Article about L1-norm and L2-norm

[http://www.chioka.in/differences-between-the-l1-norm-and-the-l2-norm-least-absolute-deviations-and-least- squares/](http://www.chioka.in/differences-between-the-l1-norm-and-the-l2-norm-least-absolute-deviations-and-least-%09squares/)

What is Standardization / Feature scaling?

What is overfitting?

Our training has focused on the particular training set so much, it has missed the poing

What is underfitting?

The model has not captured the underlying logic of the data

What is log transformation?

What is Multicollinearity?

https://statisticalhorizons.com/multicollinearity/

Logistic / Binary Regression Assumption

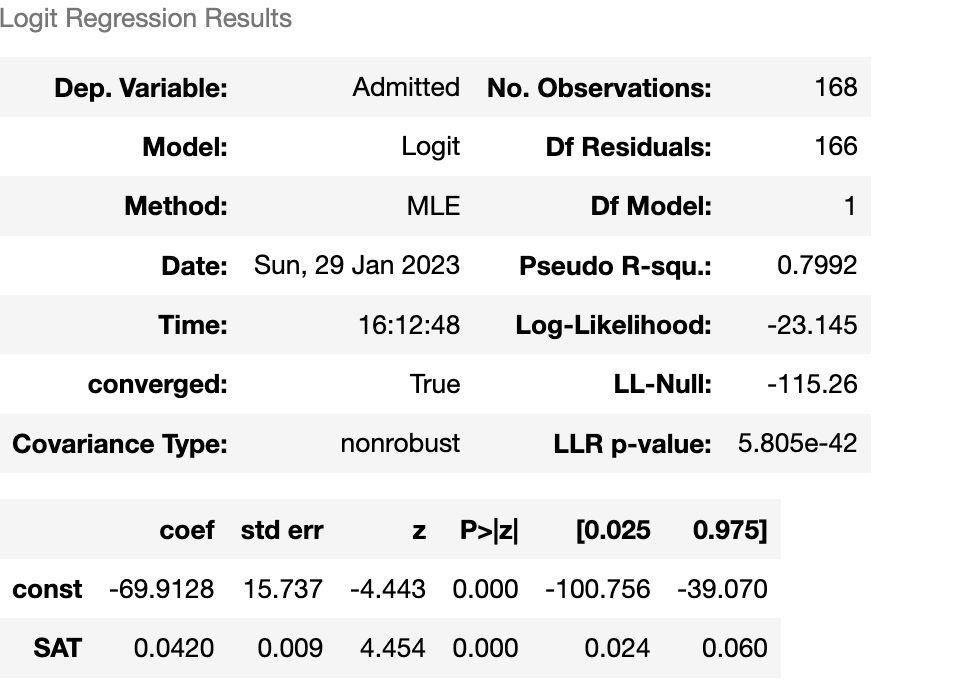
1. Linearity
2. No endogeneity
3. Normality and homoscedasticity
4. No autocorrelation
5. No multicollinearity

Logistic Regression Model p(x) = Liner Regression Model

Logit Regression Model

log( = )

log(odds) = )



MLE – Maximum likelihood estimation:

A function which estimates how likely it is that the model at hand describes the real underlying relationship of the variable

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

Log-Likelihood:

The value of the log likelihood is almost but not always negative, bigger the better

LL – Null (log likelihood -null)

The log-likelihood of a model which has no independent variables

LLR – p-value (log likelihood ratio) to check the model is significant

How to subtract two equations

From logit model - log(odds) = ), when we plot value on this for two SAT score, we get

log( = - 69.9128 +0.0420 \*

log( = - 69.9128 +0.0420 \*

subtract these two equations ----------------------------------------------------

log(log(

**Difference of 1 unit of SAT**

= => = 1.042

= 104.2 % \*

Cluster Analysis

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

Euclidean Distance:

2D space d(A,B) = d(B,A) =

3D space d(A,B) = d(B,A) =

What is Centroid?

A centroid is the mean position of a group of points (aka center of mass)

WCSS – Within Cluster Sum of Squares (value can be as low as possible)

|  |  |
| --- | --- |
| K-Means – Pros | K-Means – Cons |
| Simple to understand  Fast to cluster  Widely available  Easy to implement  Always yields a result | We need to pick K - Remedies (The Elbow Method)  Sensitive to initialization – Rem (k-means++)  Sensitive to outliers – Rem (remove outliers)  Produces spherical solutions  Standardization |