STATISTICS 8 : REGRESSION ANALYSIS WITH JUPYTER

```
# Used to create static and interactive graphical data plots
import matplotlib.pyplot as mplot

# Provides numerous math functions related to linear algebra and more
import numpy as np

# Provides numerous tools for manipulating tabular data and more
import pandas as pd

# Provides numerous tools related to statistical data analysis
import statsmodels.api as sm

# Read data from a CSV file
data = pd.read_csv('icecreamsales.csv')

# Get count, mean, standard deviation, min, max and averages for different percentiles
data.describe()

# Define the dependent variable which you want to better understand
y = data['Sales']

# Define the Y label
mplot.ylabel('Sales', fontsize=15)

# Define an independent variable
x1 = data['Temperature']

# Define the X label
mplot.xlabel('Temperature', fontsize=15)

# Create a scatter plot with the data
mplot.scatter(x1, y)

mplot.show()

# Define the intercept to the y line
x = sm.add_constant(x1)

# OLS Ordinary Least Squares : Estimates the data so a line can be drawn through
# data points
results = sm.OLS(y,x).fit()
results.summary()

mplot.scatter(x1,y)

yhat = 5.9581 * x1 + 35.5616

fig = mplot.plot(x1, yhat, lw=4, c='orange', label='regression line')

mplot.xlabel('Temperature', fontsize=15)
```

```
mplot.ylabel('Sales',fontsize=15)
mplot.show()
```

Explain the Statistics

Model : OLS : Ordinary Least Squares : One way to create a linear regression model. Minimize the dependent samples so you can estimate the unknown samples when creating a linear regression model.

Method : Least Squares : Fit data to the model by minimizing the residual samples

R-quared : Measure of how well the regression line approximates the data points. If .5 then that is a sign that half of the observed variation can be explained by the models inputs. 1 would be perfectly correlated.

Adj, R-squared : Reflects the fit of the model. Values range from 0 to 1, where higher values indicate a good fit.

F-statistic : Measures how significantly the data points fit into the regression model by measuring variation of sample means.

Prob (F-statistic) : Probability that the null hypothesis for the full model is true. Closer to zero the better the samples approach the model.

Log-Likelihood : The conditional probability that the observed data fits the model

AIC : Adjusts the log-likelihood based on the number of observations and complexity of the model. It focuses on the data points that best describe the data.

Df Residuals : Degrees of freedom of the residuals which is the difference between predicted values and the measured data.

BIC : We want a low BIC. It focuses on the shortest description of the data like AIC.

Df Model : Number of parameters in the model

Coefficient Constant : Is your Y intercept. If both dependent and independent coefficients are zero then the expected output would equal the constant coefficient.

Independent Coefficient : Represents the change of the independent variable per unit.

Standard Error : Accuracy of the coefficients

P>|t| : The P Value. A P Value less than .05 is considered statistically significant.

[.025 - .975] : Confidence Interval : Represents the range in which coefficients are likely to fall.

Omnibus : (D'Angostino's test) : Establishes whether the samples come from a normally distributed population.

Durbin-Watson : Test to see if the errors are not independent. Used to find repeating patterns that may be obstructed by noise. Its value lies between 0 and 4. If greater than 2 this is a sign

that relationships between two variables are going in opposite directions (negatively correlated). If less than 2 variables are positively correlated.

Prob(Omnibus) : Probability of Omnibus

Jarque-Bera : Tests whether the samples match a normal distribution. It never has a negative number and the further it gets from zero signals the data doesn't have a normal distribution.

Skew : Measure of the asymmetry of the probability distribution. Negative skew indicates the tail is longer on the left and the concentration of the data is on the right. Positive indicates the tail is longer on the right. 0 indicates that the tails are balanced.

Prob(JB) : The probability of Jarque-Bera

Kurtosis : Describes the shape of a probability distribution with a focus on the tails and not the peak. If the value is high that is a sign that there are more outliers. If the value is less than 3 that means there are fewer outliers. A value of 3 points towards a normal distribution. Values greater than 3 indicate more outliers.

Condition Number : Represents whether samples are highly related in our regression model. A large number indicates strong multicollinearity which means that independent variables are highly correlated with each other. This causes problems because a small number of samples are so dramatically different from others that results are corrupted.