INTERPRETION of OLS: Mithilesh Singh

Ordinary Least Squared (OLS) Regression
OLS is a linear regression Algorithm.
linear regression using the **statsmodels Python package.**

```
import statsmodels.api as sm
X = sm.add_constant(X)
```

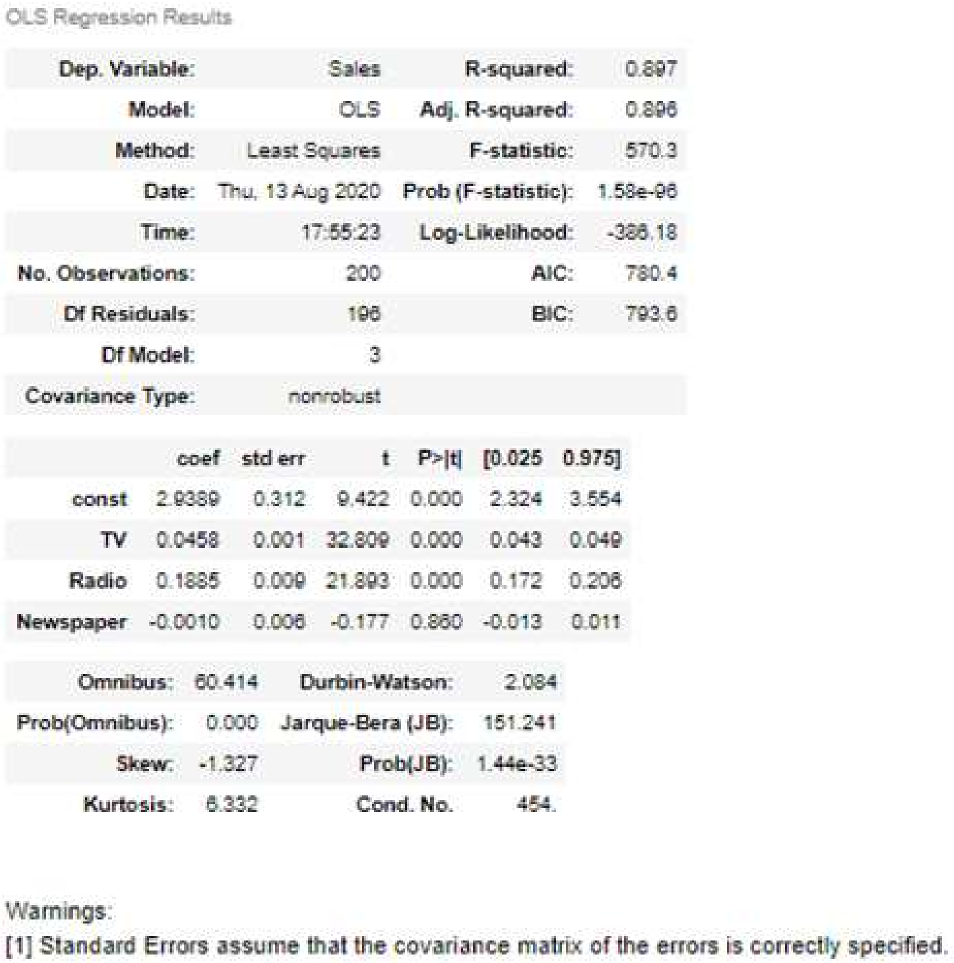
Our model needs an intercept so we add a column of 1s:

```
model = sm.OLS(y, X)
results = model.fit()
print(results.summary())
```

OLS (Ordinary Least Squared) Regression as the base model for Linear Regression.

→OLS tells more than the accuracy of the overall model.

Let's Look the Example of OLD summary: -
We will interpret each and every section of this summary table.



OLS Regression Results

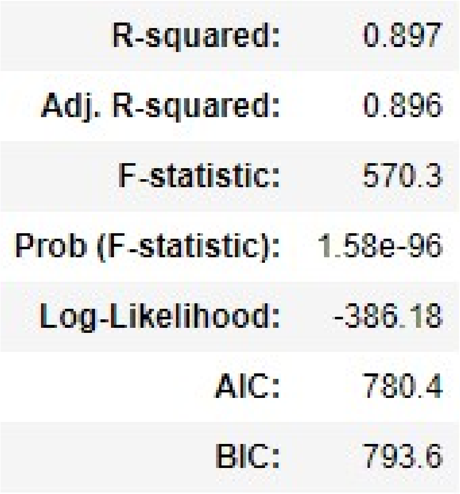
Dep. Variable:	Sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Thu, 13 Aug 2020	Prob (F-statistic):	1.58e-96
Time:	17:55:23	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

PART 1



R-squared:	0.897
Adj. R-squared:	0.896
F-statistic:	570.3
Prob (F-statistic):	1.58e-96
Log-Likelihood:	-386.18
AIC:	780.4
BIC:	793.6

R-squared: It's the degree of the variation in the dependent variable y that is explained by the dependent variables in X. for example if,R2= 89.7%. It means that our predicted values are 89.7% closer to the actual value i.e y. R2 and attain values between 0 to 1.

The drawback with an R2 score it that, more the number of variables in X, R2 has a tendency to be constant or increase even by a miniscule (very small) amount. However, the new added variable may or may not be significant.
R2 = Variance Explained by the model / Total Variance
OLS Model: Overall model R2 is 89.7%

Adjusted R-squared: This resolves the drawback of R2 score and hence is known to be more reliable. Adj. R2 doesn't consider the variables which are not significant for the model. In a single linear regression, the value of R2 and Adjusted R2 will be the same. If more number of insignificant variables are added to the model, the gap between R2 and Adjusted R2 will keep increasing.
Adjusted R Squared = 1 — [((1 — R2) * (n — 1)) / (n — k — 1)]
Where n — number of records and k is number of significant variables .
OLS Model: Adjusted R2 for the model is 89.6% which is 0.1% less than R2.

F-statistic and Prob(F-statistic): Here ANOVA is applied on the model with the following hypothesis:

H0: b1, b2, b3 (Regression coefficients) are 0 or model with no independent variables fits the data better.
H1: At least 1 of the coefficients (b1,b2,b3) is not equal to 0 or the current model with independent variable fits the data better than the intercept only model.

Now practically, having all of the independent variables to have coefficients 0 is not likely and we end up Rejecting the null hypothesis.

F-statistic = Explained variance / unexplained variance

In this OLS Model: The F-stat probability is 1.58e-96 which is much lower than 0.05 which is or alpha value. It simply means that the probability of getting at least 1 coefficient to be a nonzero value is 1.58e-96.

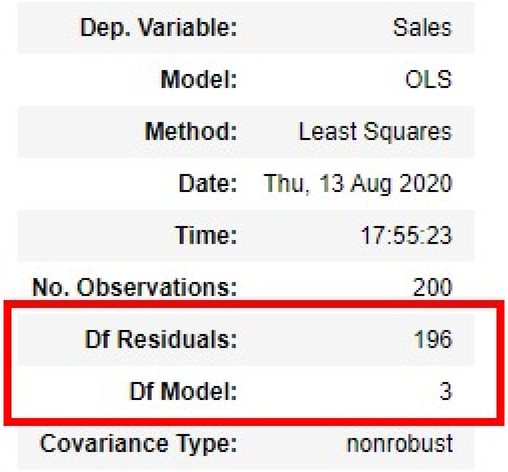
Log-Likelihood: Log Likelihood value is a measure of goodness of fit for any model or to derive the maximum likelihood estimator.
Higher the value, better is the model.
Log Likelihood can lie between -Inf to +Inf. Hence, the absolute look at the value cannot give any indication.

AIC and BIC: Akaike Information Criterion(AIC) and Bayesian Information Criterion (BIC) are 2 methods of scoring and selecting model.
AIC = -2/N * LL + 2 * k/N
BIC = -2 * LL + log(N) * k
Where N is the number of examples in the training dataset, LL is the log-likelihood of the model on the training dataset, and k is the number of parameters in the model.

The score, as defined above, is minimized, e.g. the model with the lowest AIC and BIC is selected.
The quantity calculated is different from AIC, although can be shown to be proportional to the AIC. Unlike the AIC, the BIC penalizes the model more for its complexity, meaning that more complex models will have a worse (larger) score and will, in turn, be less likely to be selected.

SECTION 2:

Over all our model is performing well with 89% accuracy.



Dep. Variable:	Sales
Model:	OLS
Method:	Least Squares
Date:	Thu, 13 Aug 2020
Time:	17:55:23
No. Observations:	200
Df Residuals:	196
Df Model:	3
Covariance Type:	nonrobust

Df Residuals:

Df here is Degrees of Freedom (DF) which indicates the number of independent values that can vary in an analysis without breaking any constraints.

Residuals in regression is simply the error rate which is not explained by the model. It’s the distance between the data point and the regression line.

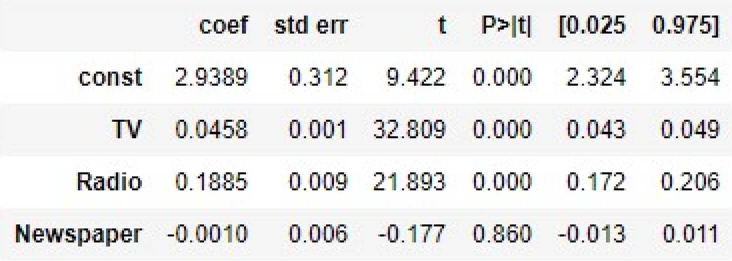
Residuals = (Observed value) — (Fitted/ Expected value)

The ***df (Residual)*** is the sample size minus the number of parameters being estimated, so it becomes df(Residual) = n — (k+1) or df(Residual) = n -k -1.
Hence the calculation in our case is:
200 (total records)-3(number of X variables) -1 (for Degree of Freedom)

Df Model: Its simple the number of X variables in the data barring the Constant variable which is 3.

SECTION 3:

central part which is the main part of the summary:

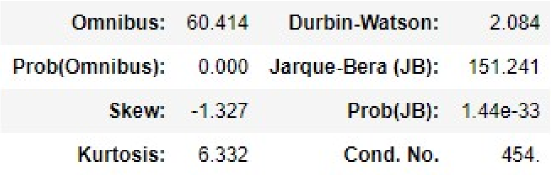


	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Central section

Now we know, the column ***coef*** is the value of b0, b1, b2 and b3. So the equation of the line is:
y = 2.94 + 0.046 * (TV) + 0.188* (Radio) + (-0.001)*(Newspaper)
Std err is the standard error for each variable, it’s the distance that the variable is away from the regression line.
***t and* P>|t|: t** is simply the t-stat value of each variable with the following hypothesis:
H0: Slope / Coefficient = 0
H1: Slope / Coefficient is not = 0
Basis this, it gives us the t stat values and the **P>|t|** gives us the p-value. With alpha at 5%, we measure if the variables are significant.
[0.025, 0.975] — At default 5% alpha or 95% Confidence interval, if the coef value lies in this region, we say that the coef value lies within the Acceptance region.
Looking at the p-values, we know we have to remove ‘Newspaper’ from our list and it’s not a significant variable. Before we come to that lets quickly interpret the last section of the model.

SECTION 4:



Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

Bottom section

Omnibus: They test whether the explained variance in a set of data is significantly greater than the unexplained variance, overall. Its a test of the skewness and kurtosis of the residual. We hope for the Omnibus score to be close to 0 and its probability close to 1 which means the residuals follow normalcy.

In our case Omnibus score is very high, way over 60 and its probability is 0. This means our residuals or error rate does not follow a normal distribution.

Skew — Its a measure of data symmetry. We want to see something close to zero, indicating the residual distribution is normal. Note that this value also drives the Omnibus.

We can see that our residuals are negatively skewed at -1.37.

Kurtosis — Its a measure of curvature of the data. Higher peaks lead to greater Kurtosis. Greater Kurtosis can be interpreted as a tighter clustering of residuals around zero, implying a better model with few outliers.

Looking at the results, our kurtosis is 6.33 which means our data doesn’t have outliers.

Durbin-Watson — The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical regression analysis. The Durbin-Watson statistic will always have a value between 0 and 4. A value of 2.0 means that there is no autocorrelation detected in the sample.

Durbin-Watson value is 2.084 which is very close to 2 and we conclude that the data doesn’t have autocorrelation.

Note: Autocorrelation, also known as serial correlation, it is the similarity between observations as a function of the time lag between them.

Jarque-Bera (JB)/Prob(JB) — JB score simply tests the normality of the residuals with the following hypothesis:

H0: Residuals follow a normal distribution

H1: Residuals don’t follow a normal distribution

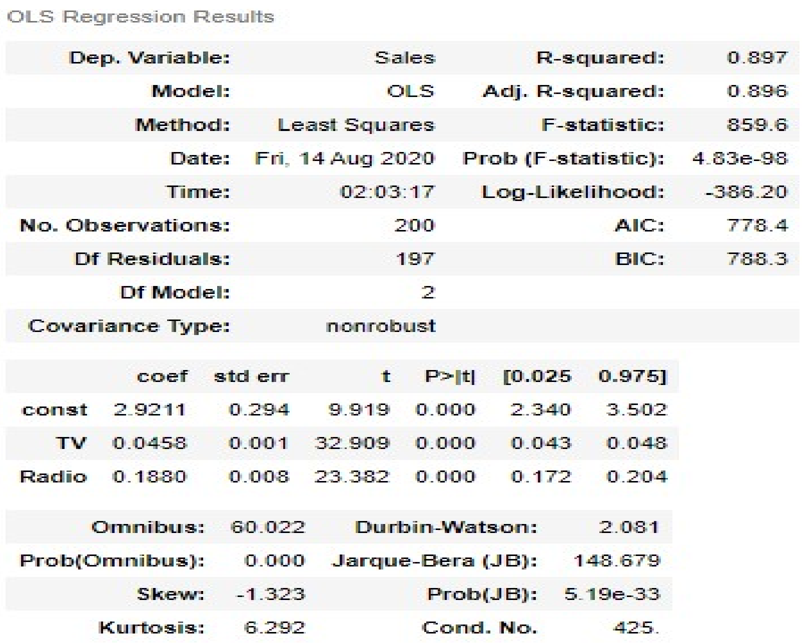
Prob(JB) is very low, close to 0 and hence we reject the null hypothesis.

Cond. No. : The condition number is used to help diagnose collinearity. Collinearity is when one independent variable is close to being a linear combination of a set of other variables.

The condition number is 454 in our case, when we reduce our variables lets see how the score reduces.

Okay we are almost at the end of our article, we have already seen the interpretation of each and every element in this OLS model. Just 1 last section where we update our OLS model and compare the results:

If we look at our model, only Newspaper with p-value 0.86 is higher than 0.05. Hence we will rebuild a model after removing the Newspaper:



```
OLS Regression Results
   Dep. Variable:            Sales           R-squared:     0.897
           Model:              OLS      Adj. R-squared:     0.896
          Method:    Least Squares         F-statistic:     859.6
            Date:  Fri, 14 Aug 2020  Prob (F-statistic):  4.83e-98
            Time:         02:03:17      Log-Likelihood:   -386.20
No. Observations:              200                 AIC:     778.4
    Df Residuals:              197                 BIC:     788.3
        Df Model:                2
 Covariance Type:        nonrobust

          coef  std err       t  P>|t|  [0.025  0.975]
const   2.9211    0.294   9.919  0.000   2.340   3.502
   TV   0.0458    0.001  32.909  0.000   0.043   0.048
Radio   0.1880    0.008  23.382  0.000   0.172   0.204

       Omnibus:  60.022     Durbin-Watson:      2.081
Prob(Omnibus):    0.000  Jarque-Bera (JB):    148.679
          Skew:  -1.323          Prob(JB):   5.19e-33
      Kurtosis:   6.292          Cond. No.       425.
```

Updated OLS model (removed Newspaper)

Please note as already mentioned, the coefficient values for each variable is dependent on the other. Hence we should always remove columns 1 by 1 so that we can gauge the difference.

When we remove the newspaper, our accuracy levels do not change however the coefficients have been updated. However, the AIC, BIC scores and Cond. No. have reduced which proves we have improved the efficiency of the model.