Md Sohel Mahmood Follow

Apr 21, 2022 · 7 min read · ✦ · Listen

Simple Explanation of Statsmodel Linear Regression Model Summary

Statsmodel library model summary explanation

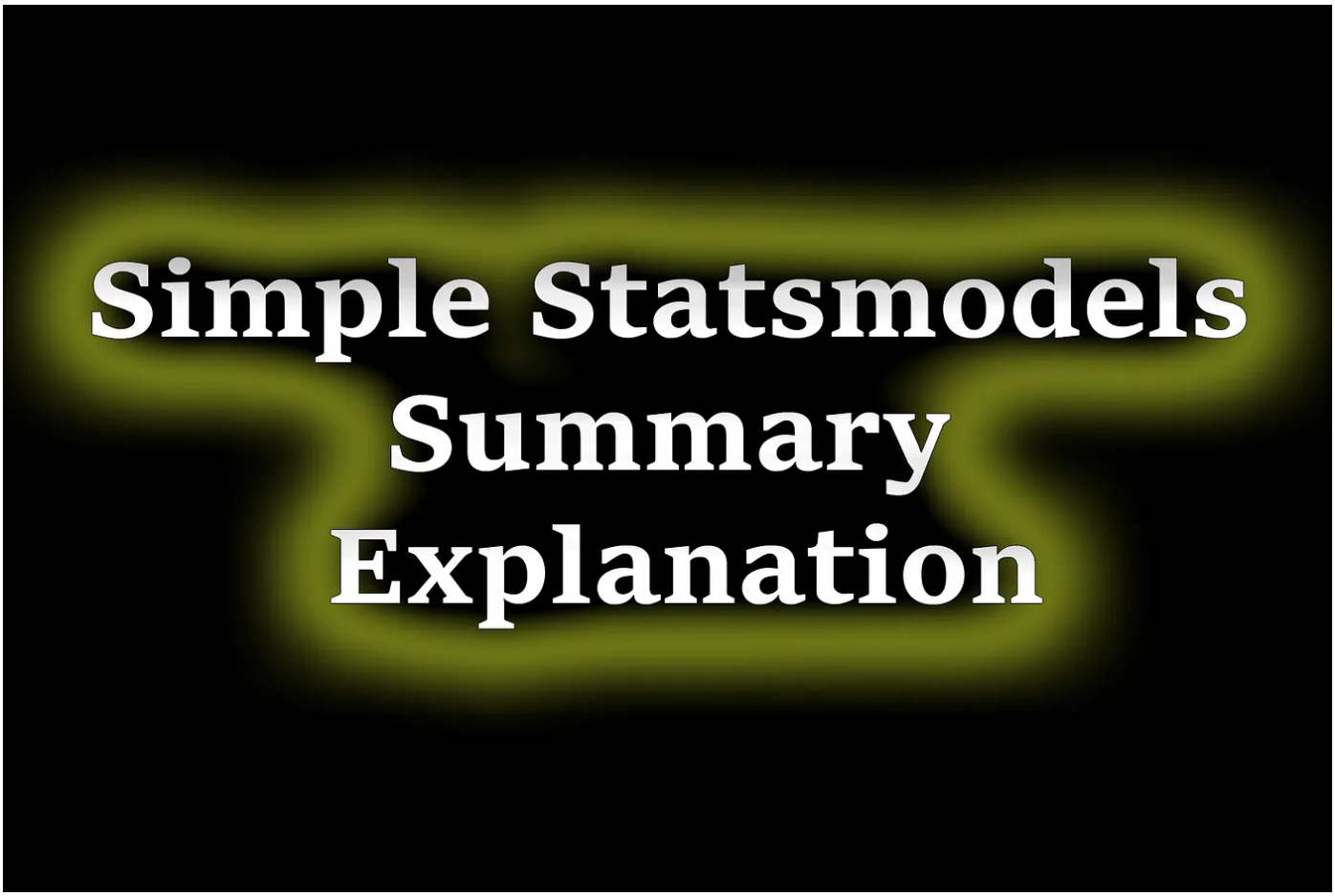


Image by Author

Introduction

Regression analysis is the bread and butter for many statisticians and data scientists. We perform simple and multiple linear regression for the purpose of prediction and always want to obtain a robust model free from any bias. In this article, I am going to discuss the summary output of python's statsmodel library using a simple example and explain a little bit how the values reflect the model performance.

Typical model summary

For the purposae of demonstration, I will use kaggle's Salary dataset (Apache 2.0 open source license). This dataset has two columns: years of experience and salary. I have two two more column: Projects and People_managing.

```
1 df = pd.read_csv("salary.csv")
2 df
```

	YearsExperience	Projects	People_managing	Salary
0	1.1	4	0	39343
1	1.3	5	0	46205
2	1.5	6	0	37731
3	2.0	3	1	43525
4	2.2	5	1	39891
5	2.9	6	1	56642
6	3.0	8	1	60150
7	3.2	7	1	54445
8	3.2	9	2	64445
9	3.7	10	2	57189
10	3.9	15	2	63218
11	4.0	12	2	55794
12	4.0	7	2	56957
13	4.1	22	3	57081

Sample data

When we use statsmodel to use all the three variables to predict Salary, we get the following summary result.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.963
Model:                            OLS   Adj. R-squared:                  0.959
Method:                 Least Squares   F-statistic:                     235.6
Date:                Thu, 21 Apr 2022   Prob (F-statistic):           1.82e-19
Time:                        16:42:36   Log-Likelihood:                -310.21
No. Observations:                  31   AIC:                             628.4
Df Residuals:                      27   BIC:                             634.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.567e+04   2221.384     11.556      0.000    2.11e+04    3.02e+04
Projects        333.4580    282.859      1.179      0.249    -246.920     913.836
People_managing -2447.4776   2426.626     -1.009      0.322   -7426.503    2531.548
YearsExperience  9633.2604   1210.014      7.961      0.000    7150.516    1.21e+04
==============================================================================
Omnibus:                        2.060   Durbin-Watson:                   1.958
Prob(Omnibus):                  0.357   Jarque-Bera (JB):                1.859
Skew:                           0.555   Prob(JB):                        0.395
Kurtosis:                       2.545   Cond. No.                         56.0
==============================================================================
```

I am going to explain all these parameters in the summary below.

Dep variable

“Salary” which is the only dependent variable in the data.

Model and Method

OLS which stands for Ordinary Least Square. The model tries to find out a linear expression for the dataset which minimizes the sum of residual squares.

DF residuals and DF model

We have total 30 observation and 4 features. Out of 4 features, 3 features are independent. DF Model is therefore 3. DF residual is calculated from total observation-

DF model-1 which is 30–3–1 = 26 in our case.

Covariance type

Covariance type is typically nonrobust which means there is no elimination of data to calculate the covariance between features. Covariance shows how two variables move with respect to each other. If this value is greater than 0, both move in same direction and if this is less than 0, the variables mode in opposite direction. Covariance is difference from correlation. Covariance does not provide the strength of the relationship, only the direction of movement whereas, correlation value is normalized and ranges between -1 to +1 and correlation provides the strength of relationship. If we want to obtain robust covariance, we can declare cov_type=HC0/HC1/HC2/HC3. However, the statsmodel documentation is not that rich to explain all these. HC stands for heteroscedasticity consistent and HC0 implements the simplest version among all.

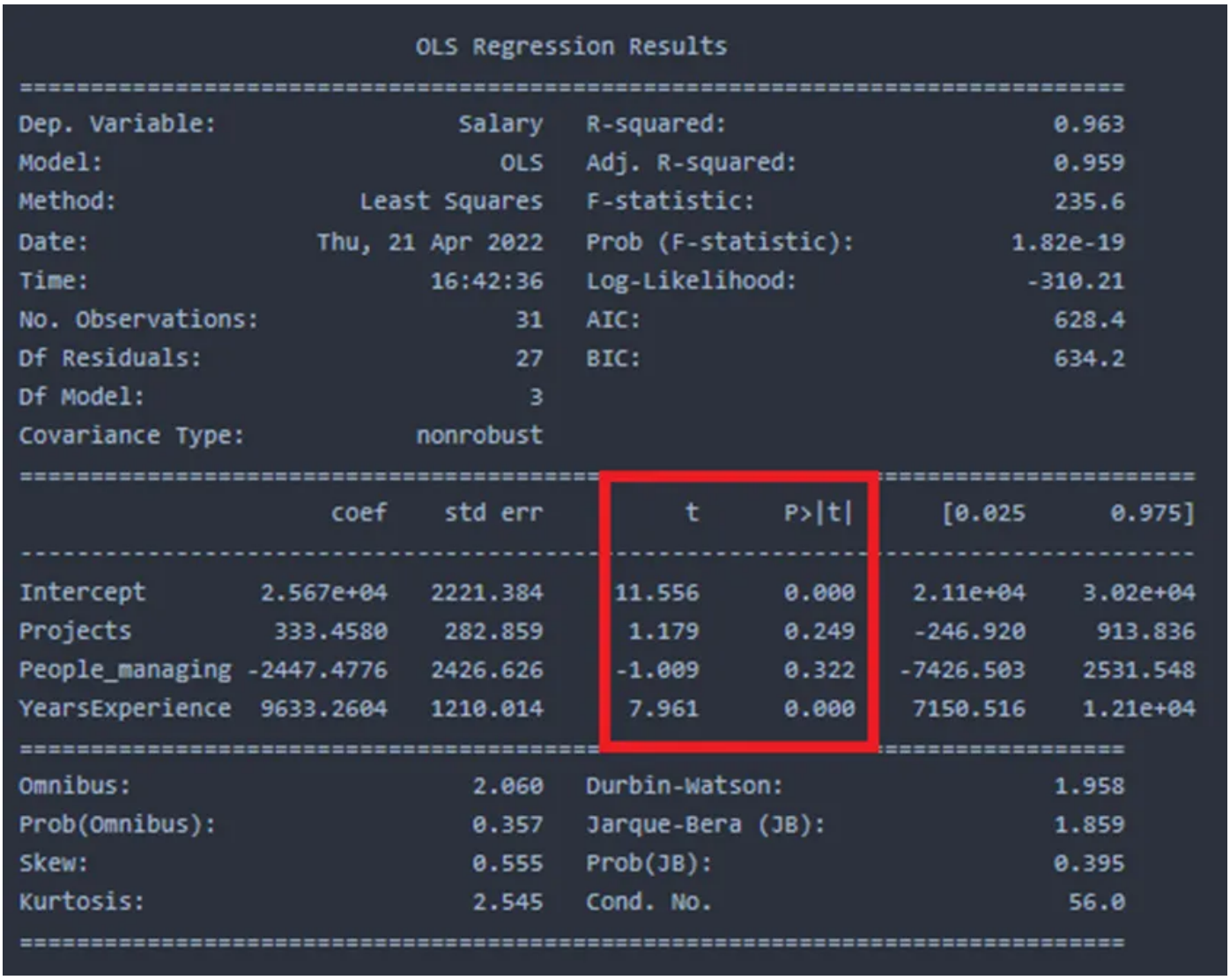
R-squared

R-squared value is the coefficient of determination which indicates the percentage of the variability if the data explained by the selected independent variables.

Adj. R-squared

As we add more and more independent variables to our model, the R-squared values increases but in reality, those variables do not necessarily make any contribution towards explaining the dependent variable. Therefore addition of each unnecessary variables needs some sort of penalty. The original R-squared values is adjusted when there are multiple variables incorporated. In essence, we should always look for adjusted R-squared value while performing multiple linear regression. For a single independent variable, both R-squared and adjusted R-squared value are same.

Before moving to F-statistics, we need to understand the t-statistics first. T-statistics are provided in the table shown below.



OLS Regression Results

Dep. Variable:	Salary	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.959
Method:	Least Squares	F-statistic:	235.6
Date:	Thu, 21 Apr 2022	Prob (F-statistic):	1.82e-19
Time:	16:42:36	Log-Likelihood:	-310.21
No. Observations:	31	AIC:	628.4
Df Residuals:	27	BIC:	634.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.567e+04	2221.384	11.556	0.000	2.11e+04	3.02e+04
Projects	333.4580	282.859	1.179	0.249	-246.920	913.836
People_managing	-2447.4776	2426.626	-1.009	0.322	-7426.503	2531.548
YearsExperience	9633.2604	1210.014	7.961	0.000	7150.516	1.21e+04

Omnibus:	2.060	Durbin-Watson:	1.958
Prob(Omnibus):	0.357	Jarque-Bera (JB):	1.859
Skew:	0.555	Prob(JB):	0.395
Kurtosis:	2.545	Cond. No.	56.0

> *coef and std err*

The coef column represents the coefficients for each independent variable along with intercept value. Std err is the standard deviation of the corresponding variable's coefficient across all the data points. When using only one predicting variable, the standard error can be obtained from this two dimensional space as shown below

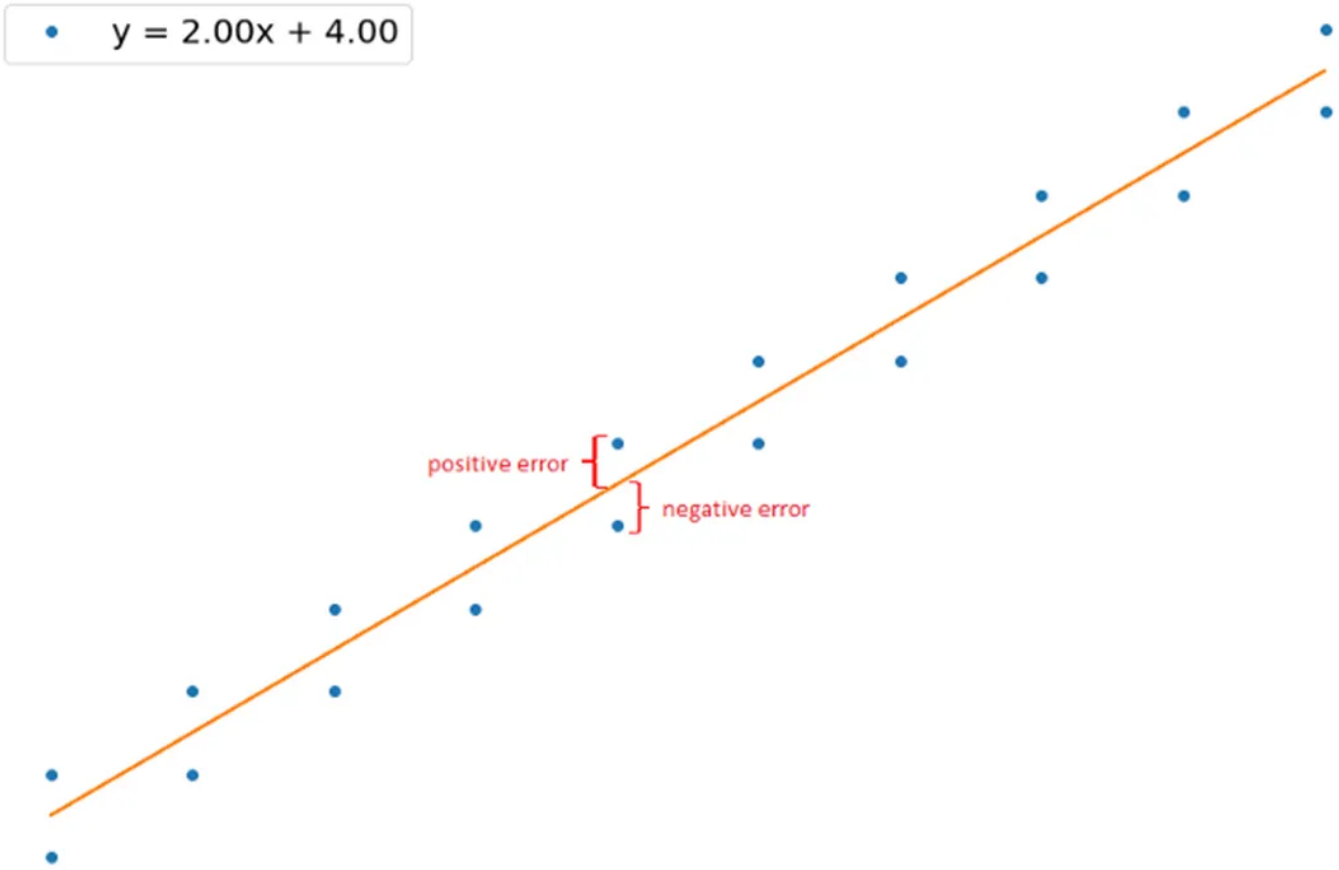
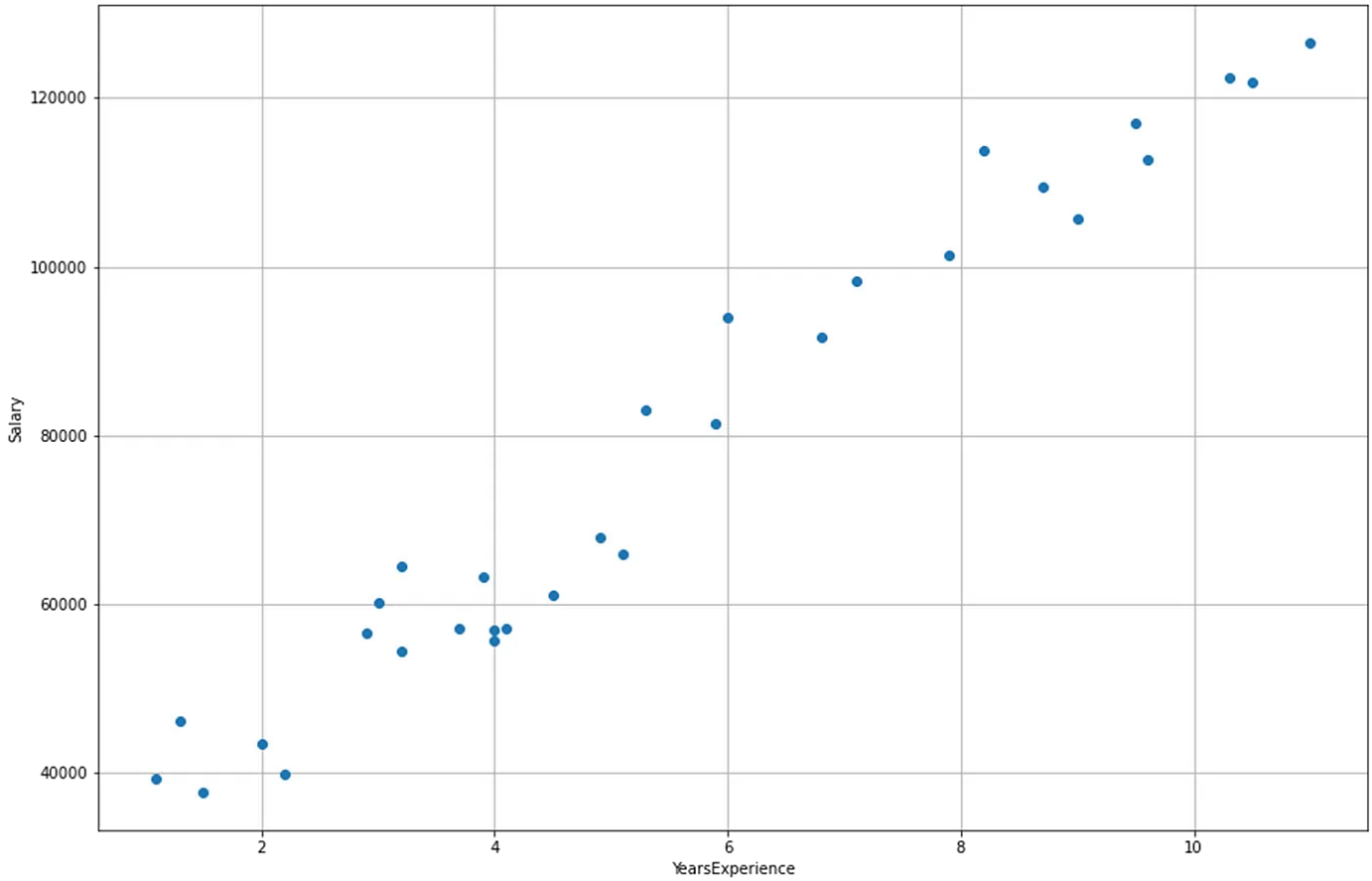


Image by Author

t-values and P>|t|

The t-column provides the t-values corresponding to to each independent variables. For example here Projects, People_managing and Salary all have different t-values as well as different p-values associated with each variables. T-statistics are used to calculate the p-values. Typically when p-value is less than 0.05, it indicates a strong evidence against null hypothesis which states that the corresponding independent variable has no effect on the dependent variable. P-value of 0.249 for Projects says us that there is 24.9% chance that Projects variables has no effect on Salary. It seems YearsExperience got 0 p-value indicating that the data for YearsExperience is statistically significant since is is less than the critical limit (0.05). In this case, we can reject the null hypothesis and say that YearsExperience data is significantly controlling the Salary.



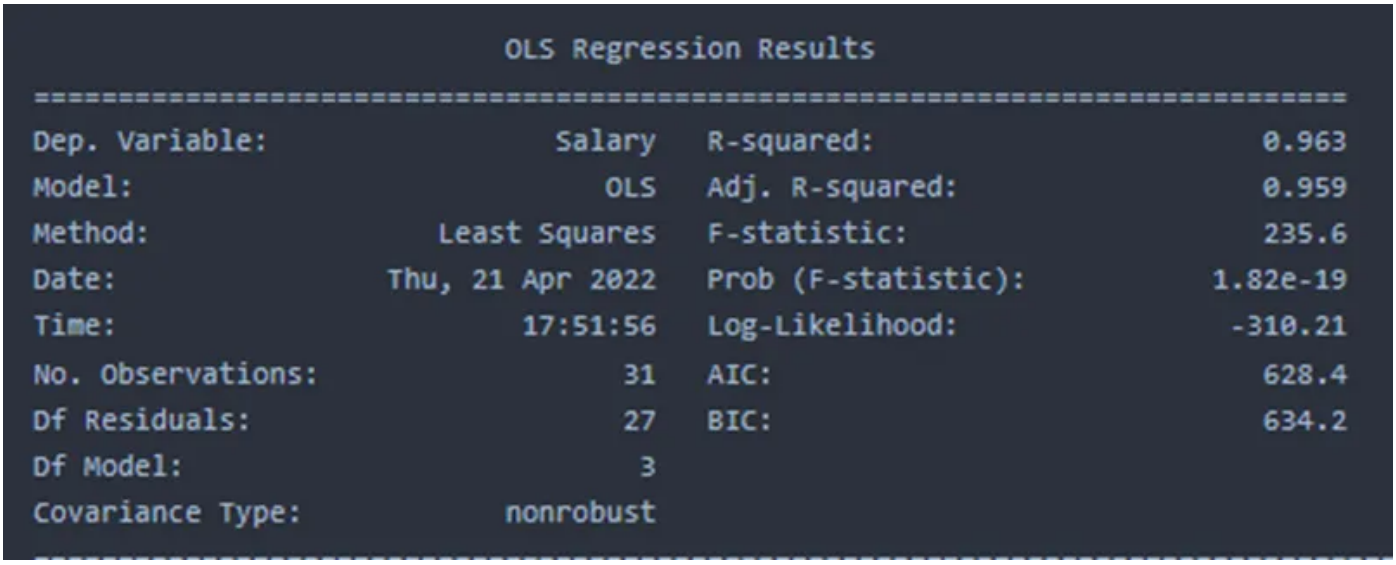
Years of Experience against Salary showing strong correlation

F-statistics

F-test provides a way to check all the independent variables all together if any of those are related to the dependent variable. If Prob(F-statistic) is greater than 0.05, there is no evidence of relationship between any of the independent variable with the output. If it is less than 0.05, we can say that there is at least one variable which is significantly related with the output. In our example, the p-value is less than 0.05 and therefore, one or more than one of the independent variable are related to output variable Salary. We have seen previously that YearsExperience is significantly related with Salary but others are not. Therefore, the F-test data supports the t-test outcomes. However, there may be some cases when prob(F-statistic) may be greater than 0.05 but one of the independent variable shows strong correlation. This is because each t-test is carried out with different set of data whereas F-test checks the combined effect including all variables globally.

Log-likelihood

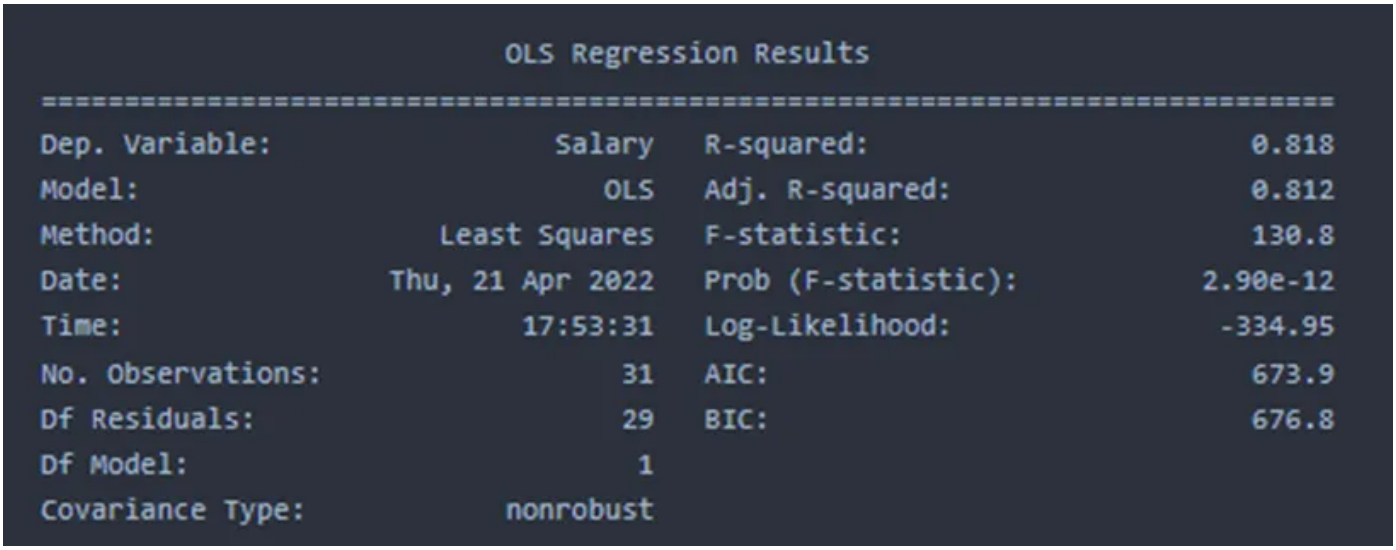
The log-likelihood value is a measure for fit of the model with the given data. It is useful when we compare two or more models. The higher the value of log-likelihood, the better the model fits the given data. It can range from negative infinity to positive infinity.



OLS Regression Results

Dep. Variable:	Salary	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.959
Method:	Least Squares	F-statistic:	235.6
Date:	Thu, 21 Apr 2022	Prob (F-statistic):	1.82e-19
Time:	17:51:56	Log-Likelihood:	-310.21
No. Observations:	31	AIC:	628.4
Df Residuals:	27	BIC:	634.2
Df Model:	3		
Covariance Type:	nonrobust		

log-likelihood when all three variables are included



OLS Regression Results

Dep. Variable:	Salary	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.812
Method:	Least Squares	F-statistic:	130.8
Date:	Thu, 21 Apr 2022	Prob (F-statistic):	2.90e-12
Time:	17:53:31	Log-Likelihood:	-334.95
No. Observations:	31	AIC:	673.9
Df Residuals:	29	BIC:	676.8
Df Model:	1		
Covariance Type:	nonrobust		

log-likelihood when only “Projects” is included

When all three independent variables are incorporated in the model, the log-likelihood value is -310.21 which is higher than -334.95 when only Projects data is included. This mean the first model fits the data better. It also goes hand in hand with R-squared values as seen above.

AIC and BIC

AIC (stands for Akaike's Information Criteria developed by Japanese statistician Hirotugo Akaike) and BIC (stands for Bayesian Information Criteria) are also used as criteria for model robustness. The goal is to minimize these values to get a better model. I have another article where I have discussed on these topics.

Simple Stepwise and Weighted Regression Model

Stepwise and Weighted Regression

Model Stepwise and Weighted Regressiontowardsdatascience.com

Omnibus and Prob(Omnibus)

Omnibus test checks the normality of the residuals once the model is deployed. If the value is zero, it means the residuals are perfectly normal. Here, in the example prob(Omnibus) is 0.357 indicating that there is 35.7% chance that the residuals the normally distributed. For a model to be robust, besides checking R-squared and other rubrics, the residual distribution is also required to be normal ideally. In other words, the residual should not follow any pattern when plotted against the fitted values.

Skew and Kurtosis

Skew values tells us the skewness of the residual distribution. Normally distributed variables have 0 skew values. Kurtosis is a measure of light-tailed or heavy-tailed distribution compared to normal distribution. High kurtosis indicates the distribution is too narrow and low kurtosis indicates the distribution is too flat. A kurtosis value between -2 and +2 is good to prove normalcy.

Durbin-Watson

Durbin-Watson statistic provides a measure of autocorrelation in the residual. If the residual values are autocorrelated, the model becomes biased and it is not expected. This simply means that one value should not be depending on any of the previous values. An ideal value for this test ranges from 0 to 4.

Jarque-Bera (JB) and Prob(JB)

Jarque-Bera (JB) and Prob(JB) is similar to Omni test measuring the normalcy of the residuals.

Condition Number

High condition number indicates that there are possible multicollinearity present in the dataset. If only one variable is used as predictor, this value is low and can be ignored. We can proceed like stepwise regression and see if there is any multicollinearity added when additional variables are included.

Conclusion

We have discussed all the summary parameters from statsmodel output. This will useful for readers who are interested to check all the rubrics for a robust model., Most of the time, we look for R-squared value to make sure that the model explains most of the variability but we have seen that there is much more than that.

Thanks for reading

Github page

Youtube Channel

Multiple Linearregression Simple Linear Regression Statsmodels Python

Statistics