Data science and Machine learning with Python

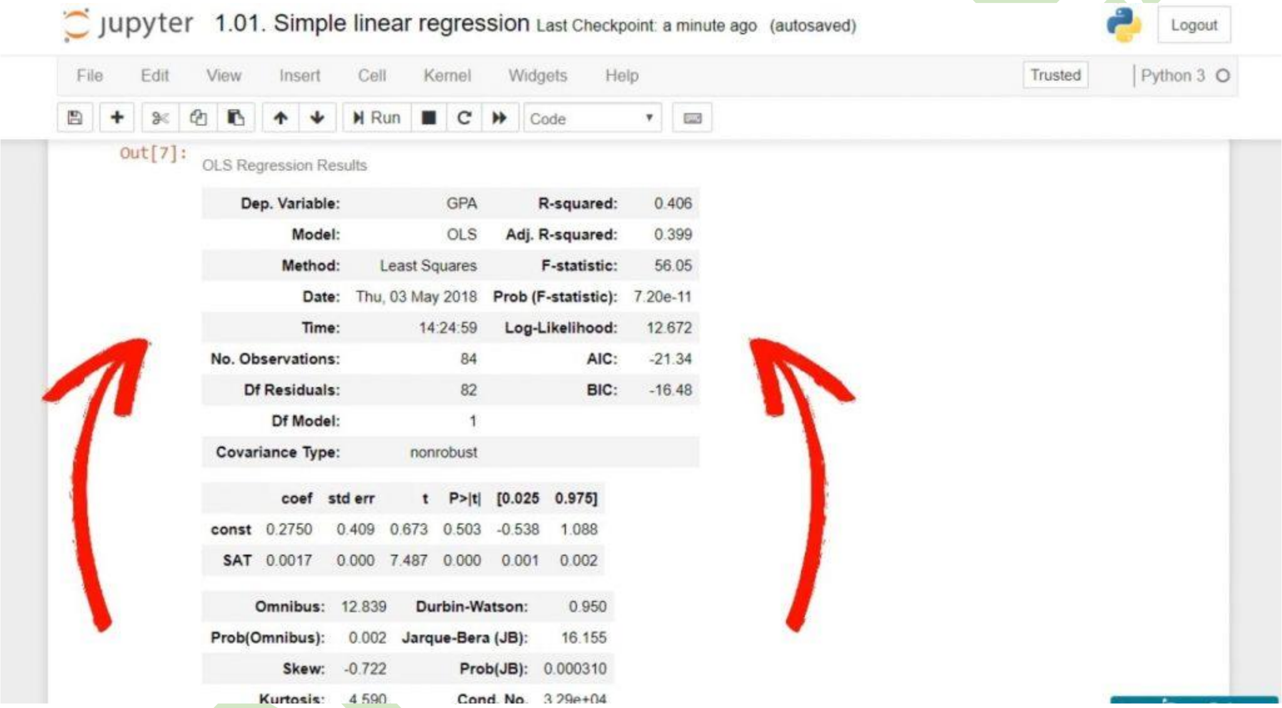
Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

Linear regression (continued)

How to Interpret the Regression Table

Now, let’s figure out how to interpret the **regression table** we saw earlier in our **linear regression** example.

While the graphs we have seen so far are nice and easy to understand. When you perform **regression analysis**, you’ll find something different than a **scatter plot** with a **regression line**. The graph is a visual representation, and what we really want is the equation of the model, and a measure of its significance and explanatory power. This is why the **regression** summary consists of a few tables, instead of a graph.



Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

Omnibus:	12.839	Durbin-Watson:	0.950
Prob(Omnibus):	0.002	Jarque-Bera (JB):	16.155
Skew:	-0.722	Prob(JB):	0.000310
Kurtosis:	4.590	Cond. No.	3.29e+04

Let’s find out how to read and understand these tables.

Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

The 3 Main Tables

Typically, when using *statsmodels*, we'll have three main tables – a *model summary*

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Thu, 03 May 2018	**Prob (F-statistic):**	7.20e-11
Time:	14:24:59	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

a *coefficients table*

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Thu, 03 May 2018	**Prob (F-statistic):**	7.20e-11
Time:	14:24:59	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

and some additional tests.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Thu, 03 May 2018	**Prob (F-statistic):**	7.20e-11
Time:	14:24:59	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

Omnibus:	12.839	**Durbin-Watson:**	0.950
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	16.155
Skew:	-0.722	**Prob(JB):**	0.000310
Kurtosis:	4.590	**Cond. No.**	3.29e+04

Certainly, these tables contain a lot of information, but we will focus on the most important parts.

We will start with the *coefficients table*.

The Coefficients Table

We can see the coefficient of the intercept, or the constant as they've named it in our case.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

Both terms are used interchangeably. In any case, it is 0.275, which means b_0 is 0.275.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Thu, 03 May 2018	**Prob (F-statistic):**	7.20e-11
Time:	14:24:59	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

$$\hat{y} = \underbrace{b_0}_{0.275} + b_1 x_1$$

Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

Looking below it, we notice the other coefficient is 0.0017. This is our b_1. These are the only two numbers we need to define the **regression equation**.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

$$\hat{y} = \underbrace{b_0}_{0.275} + \underbrace{b_1}_{0.0017} x_1$$

Therefore,

$\hat{y} = 0.275 + 0.0017 * x1$.

Or GPA equals 0.275 plus 0.0017 times SAT score.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

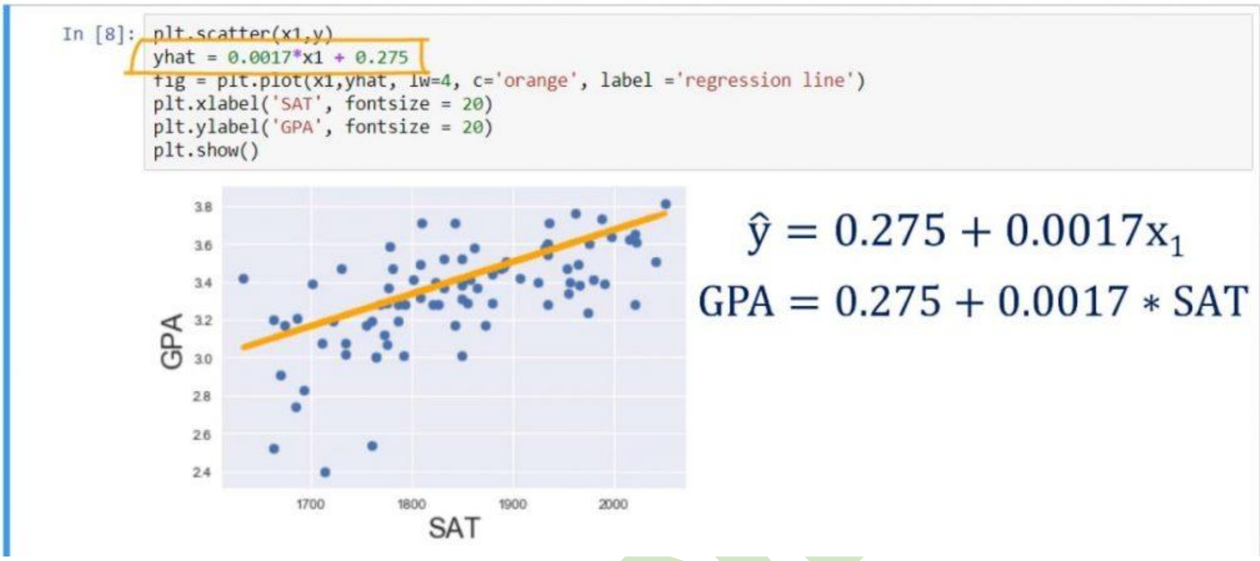
$$\hat{y} = 0.275 + 0.0017 x_1$$

$$GPA = 0.275 + 0.0017 * SAT$$

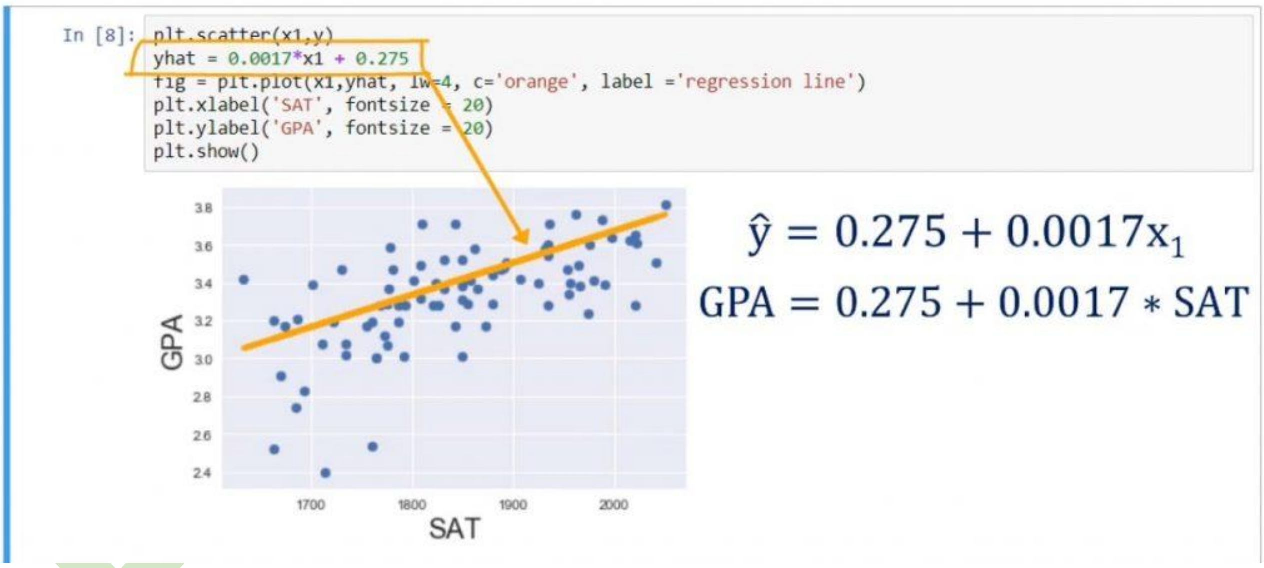
So, this is how we obtain the **regression equation**.

A Quick Recap

Let’s take a step back and look at the code where we plotted the **regression line**. We have plotted the **scatter plot** of SAT and GPA. That’s clear. After that, we created a variable called: y hat($\hat{y}$). Moreover, we imported the *seaborn* library as a ‘skin’ for *matplotlib*. We did that in order to display the regression in a prettier way.

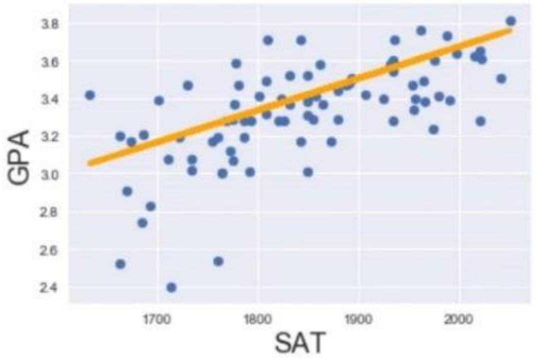


That’s the **regression line** – the predicted variables based on the data.



Finally, we plot that line using the plot method.

```
In [8]: plt.scatter(x1,y)
        yhat = 0.0017*x1 + 0.275
        fig = plt.plot(x1,yhat, lw=4, c='orange', label ='regression line')
        plt.xlabel('SAT', fontsize = 20)
        plt.ylabel('GPA', fontsize = 20)
        plt.show()
```



$$\hat{y} = 0.275 + 0.0017x_1$$

$$\text{GPA} = 0.275 + 0.0017 * \text{SAT}$$

Naturally, we picked the coefficients from the **coefficients table** – we didn't make them up.

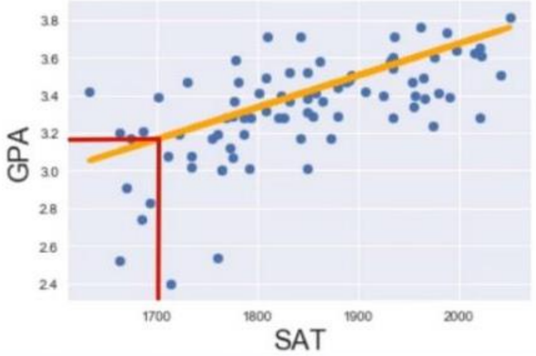
The Predictive Power of Linear Regressions

You might be wondering if that prediction is useful.

Well, knowing that a person has scored 1700 on the SAT, we can substitute in the equation and obtain the following:

0.275 + 0.0017 * 1700, which equals 3.165. So, the expected GPA for this student, according to our model is 3.165.

```
In [8]: plt.scatter(x1,y)
        yhat = 0.0017*x1 + 0.275
        fig = plt.plot(x1,yhat, lw=4, c='orange', label ='regression line')
        plt.xlabel('SAT', fontsize = 20)
        plt.ylabel('GPA', fontsize = 20)
        plt.show()
```



$$\hat{y} = 0.275 + 0.0017x_1$$

$$GPA = 0.275 + 0.0017 * SAT$$

$$3.165 = 0.275 + 0.0017 * 1700$$

And that's the predictive power of **linear regressions** in a nutshell!

The Standard Errors

What about the other cells in the table?

The **standard errors** show the accuracy of prediction for each variable. The lower the **standard error**, the better the estimate!

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

The T-Statistic

The next two values are a T-statistic and its P-value.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

If you have gone over our other tutorials, you may know that there is a **hypothesis** involved here. The **null hypothesis** of this test is: $\beta = 0$. In other words, is the coefficient equal to zero?

Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

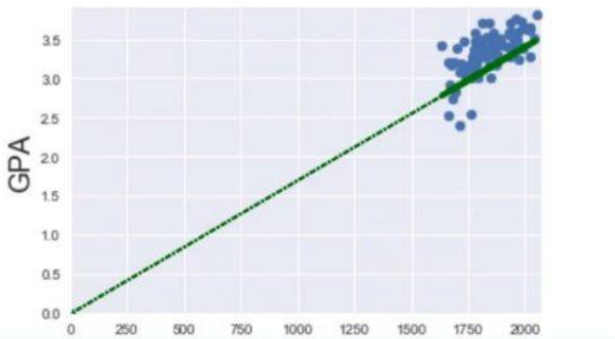
H_0: $\beta=0$

Is the coefficient equal to zero?

The Null Hypothesis

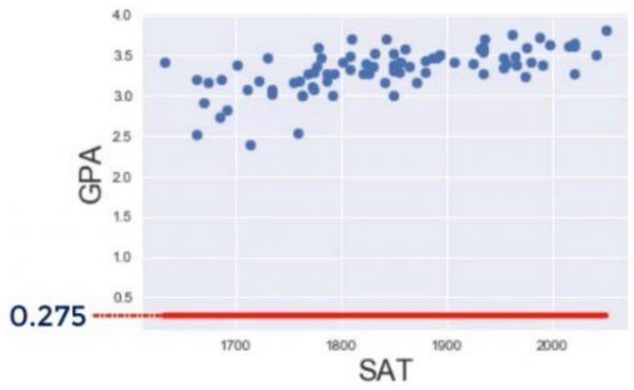
If a coefficient is zero for the intercept(b_0), then the line crosses the y-axis at the origin. You can get a better understanding of what we are talking about, from the picture below.

In [11]:
```
plt.scatter(x1,y)
yhat = 0.0017*x1 + 0
fig = plt.plot(x1,yhat, lw=4, c='green', label ='regression line')
plt.xlabel('SAT', fontsize = 20)
plt.ylabel('GPA', fontsize = 20)
plt.xlim(0)
plt.ylim(0)
plt.show()
```



If $\beta_1$is zero, then 0 * x will always be 0 for any x, so this variable will not be considered for the model. Graphically, that would mean that the regression line is horizontal – always going through the intercept value.

Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.



$$\text{If } b_1 = 0, \text{ then } \hat{y} = b_0$$

The P-Value

Let's paraphrase this test. Essentially, it asks, is this a useful variable? Does it help us explain the variability we have in this case? The answer is contained in the *P-value* column.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

As you may know, a *P-value* below 0.05 means that the variable is significant. Therefore, the coefficient is most probably different from 0. Moreover, we are longing to see those three zeroes.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

What does this mean for our **linear regression** example?

Well, it simply tells us that SAT score is a significant variable when predicting college GPA.

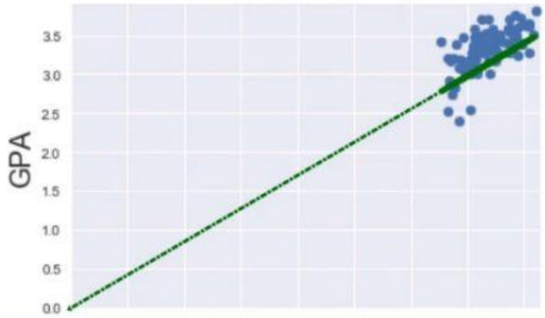
What you may notice is that the intercept *p-value* is not zero.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

Let's think about this. Does it matter that much? This test is asking the question: Graphically, that would mean that the **regression line** passes through the origin of the graph.

Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

```
In [11]: plt.scatter(x1,y)
yhat = 0.0017*x1 + 0
fig = plt.plot(x1,yhat, lw=4, c='green', label ='regression line')
plt.xlabel('SAT', fontsize = 20)
plt.ylabel('GPA', fontsize = 20)
plt.xlim(0)
plt.ylim(0)
plt.show()
```



Usually, this is not essential, as it is causal relationship of the *X*s we are interested in.

The F-statistic

The last measure we will discuss is the F-statistic. We will explain its essence and see how it can be useful to us.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

Omnibus:	12.839	Durbin-Watson:	0.950
Prob(Omnibus):	0.002	Jarque-Bera (JB):	16.155
Skew:	-0.722	Prob(JB):	0.000310
Kurtosis:	4.590	Cond. No.	3.29e+04

Much like the Z-statistic which follows a **normal distribution** and the T-statistic that follows a **Student's T distribution**, the F-statistic follows an **F distribution**.

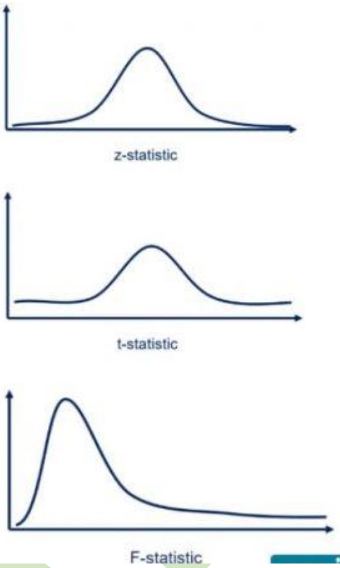
Out[7]:

OLS Regression Results

Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

Omnibus:	12.839	Durbin-Watson:	0.950
Prob(Omnibus):	0.002	Jarque-Bera (JB):	16.155
Skew:	-0.722	Prob(JB):	0.000310



We are calling it a statistic, which means that it is used for tests. The test is known as the test for overall significance of the model.

The Null Hypothesis and the Alternative Hypothesis

The **null hypothesis** is: all the βs are equal to zero simultaneously.

The **alternative hypothesis** is: at least one β differs from zero.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

F-test:

$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$

$H_1:$ at least one $\beta_i \neq 0$

This is the interpretation: if all βs are zero, then none of the *independent variables* matter. Therefore, our model has no merit.

In our case, the F-statistic is 56.05.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Thu, 03 May 2018	**Prob (F-statistic):**	7.20e-11
Time:	14:24:59	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

The cell below is its *P-value*.

Out[7]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.406	
Model:	OLS	**Adj. R-squared:**	0.399	
Method:	Least Squares	**F-statistic:**	56.05	
Date:	Thu, 03 May 2018	**Prob (F-statistic):**	7.20e-11	~ 0.000
Time:	14:24:59	**Log-Likelihood:**	12.672	
No. Observations:	84	**AIC:**	-21.34	
Df Residuals:	82	**BIC:**	-16.48	
Df Model:	1			
Covariance Type:	nonrobust			

As you can see, the number is really low – it is virtually 0.000. We say the overall model is significant.

Important: Notice how the *P-value* is a universal measure for all tests. There is an F-table used for the F-statistic, but we don't need it, because the *P-value* notion is so powerful.

The F-test is important for **regressions**, as it gives us some important insights. Remember, the lower the F-statistic, the closer to a non-significant model.

Moreover, don't forget to look for the three zeroes after the dot!

What We Learned

Well, that was a long journey, wasn't it? We embarked on it by first learning about what a **linear regression** is. Then, we went over the process of creating one. We also went over a **linear regression** example. Afterwards, we talked about the **simple linear regression** where we introduced the **linear regression equation**. By then, we were done with the theory and got our hands on the keyboard and explored another **linear regression** example in Python! We imported the relevant libraries and loaded the data. We cleared up when exactly we need to create **regressions** and started creating our own. The process consisted of several steps which, now, you should be able to perform with ease. Afterwards, we began interpreting the **regression table**. We mainly discussed the coefficients table. Lastly, we explained why the F-statistic is so important for **regressions**.

LEARN AND LEAD