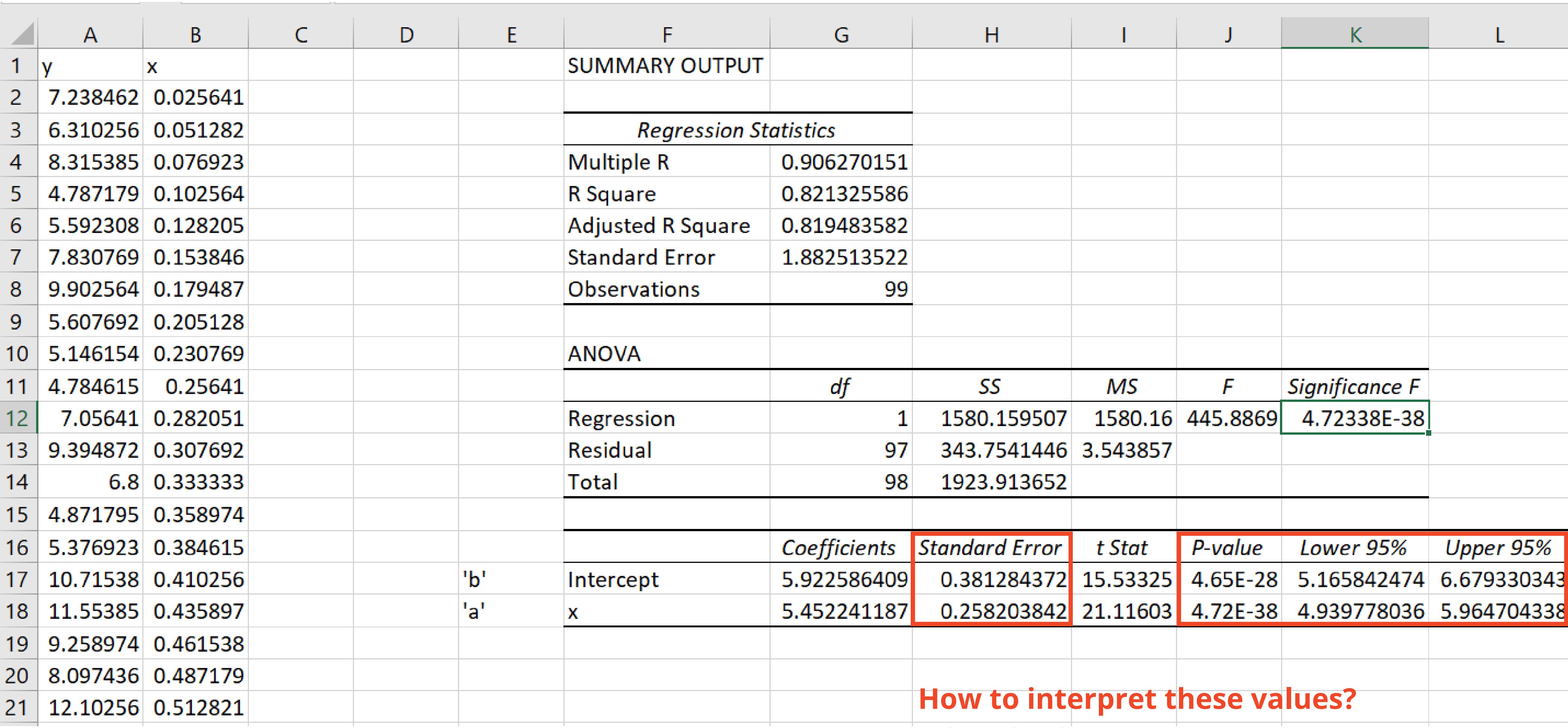


	A	B	C	D	E	F	G	H	I	J	K	L
1	y	x				SUMMARY OUTPUT						
2	7.238462	0.025641										
3	6.310256	0.051282				*Regression Statistics*						
4	8.315385	0.076923				Multiple R	0.906270151					
5	4.787179	0.102564				R Square	0.821325586					
6	5.592308	0.128205				Adjusted R Square	0.819483582					
7	7.830769	0.153846				Standard Error	1.882513522					
8	9.902564	0.179487				Observations	99					
9	5.607692	0.205128										
10	5.146154	0.230769				ANOVA						
11	4.784615	0.25641					*df*	*SS*	*MS*	*F*	*Significance F*	
12	7.05641	0.282051				Regression	1	1580.159507	1580.16	445.8869	4.72338E-38	
13	9.394872	0.307692				Residual	97	343.7541446	3.543857			
14	6.8	0.333333				Total	98	1923.913652				
15	4.871795	0.358974										
16	5.376923	0.384615					*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	10.71538	0.410256			'b'	Intercept	5.922586409	0.381284372	15.53325	4.65E-28	5.165842474	6.679330343
18	11.55385	0.435897			'a'	x	5.452241187	0.258203842	21.11603	4.72E-38	4.939778036	5.964704338
19	9.258974	0.461538										
20	8.097436	0.487179										
21	12.10256	0.512821										

How to interpret these values?

What do they mean?

The problem we are trying to solve:

- We have a dataset (x,y) and we know that it is a **random**, yet **representative sample** (let's call it sample **s1**) of the population
- Based on **s1** we have fitted an SLR model, **y = 5.4522 * x + 5.9226** by **estimating** the coefficients **a1=5.4522, b1=5.9226** (by minimizing SSE)
- If we had a different random sample **s2**, the calculated values **a2, b2** would have been different. In general, a random sample **si** will result in coefficient values **ai, bi** ...
- So, statistically, the calculated coefficient **ai** can possibly assume different values, depending on the **random sample si** that we get for analysis
- There is an important question that needs to be answered: **What is the probability that the calculated value of ai will be close to or equal to ZERO for some of the si?**
- Why is this a critical question?
    - In the regression model **y = a * x + b**, If **a** is **ZERO** then we do not really have a regression model - ie. y cannot really be predicted in terms of x !
- In the context of our example, the calculated value of **a = 5.4522 ..**
    - What is the probability that this value is obtained **by chance**, due to the peculiarity of sample **s1**?
    - What is the probability that other random samples, **si**, will result in values of **ai** that are very close or equal to **ZERO** - thereby making the model invalid?
    - (*BTW, why only **a**? Why are we not much concerned about **b**?*)
- Can we prove, based only on **s1**, that for any other **si** the calculated value of **ai** has a very high probability of being closer to **5.4522**, and almost never close to **ZERO - thereby establishing the validity of the model?**
    - More practically, can we prove, based only on **s1**, that for any other si the calculated value of **ai** has more than 95% probability of being closer to 5.4522, and less than 5% probability of being close to ZERO? (BTW, *why are we talking about 95% and 5%, and not 100% and 0%?*)
    - So, finally, it comes down to whether we can predict the **spread** of the values of **ai** based on just **s1** and **a1**
    - The **Sampling Distribution of *coefficient 'a'*** and the **Central Limit Theorem** help us to calculate this probability and, thus, decide about the validity of the **s1** based Regression model.

We will establish the **sampling distribution** of coefficients **a**,**b** based on s1, a1, b1 as follows:

- Based on **s1** calculate the coefficients **a1 (= 5.4522)** and **b1 (= 5.9226) -** both these are ***statistics*** and it is our goal to check if these values are obtained by chance, or they truly represent the values ***obtainable*** from most other samples as well.
    - The Sampling Distribution of any *statistic* provides us the mechanism to make this assessment
- To establish the Sampling Distributions of coefficients **a** and **b** we need to find out their standard deviations - the **Standard Errors**. In this context, note the following:
    - Our earlier discussion, we covered the **Sampling Distribution of the Sample Mean**. However, now our object of study is **not** the sample mean but the coefficients **a1**, **b1** which are calculated from **s1** using optimization methods. Hence the earlier standard error formula (= sigma/sqrt(n)) does not hold in this case.
- The Standard Error related to coefficients a1 and b1 are given by the formulae:

$$SE(a) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad SE(b) = \sqrt{\frac{\hat{\sigma}^2}{n}\left(1+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$$

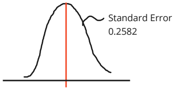
$\hat{\sigma}^2$ is the estimated variance of the error term in the model.

- On plugging the sample data (**s1**) into these formulae, the Standard Errors of the sampling distribution of **a** and **b** are 0.2582 and 0.3813, respectively, and we know that both Sampling Distributions follow the Normal Distribution.

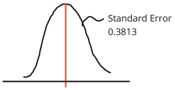
In the formulae alongside we need the value of 'sigma' the population standard deviation. Wherefrom do we get it?

We rely on our sample 's1' for that!

Recollect that s1 is supposed to be representative of the population. Hence, we can assume that the standard deviation of the sample is close to the standard deviation of the population. As we are using this value for estimating the prediction error of a and b, this assumption is acceptable.

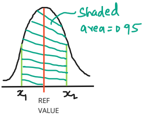


Sampling Distribution of coefficient **a**

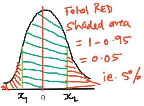


Sampling Distribution of coefficient **b**

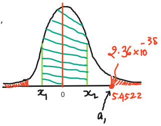
- Lets again recollect that the Sampling Distributions reflect the variations in the *statistics* calculated using different random samples **si**.
- Lets also recollect that the area under the Sampling Distribution curve (formally known as the **Probability Density Function**) represents probability. For example, the shaded area in the figure alongside gives the probability of the values between x1 and x2 - also note that the **total area** under the probability distribution curve is 1.0
- For any REF VALUE (we will use ZERO - why?), and the above calculated STANDARD ERROR values, we can find out the limits x1 and x2 using Normal Distribution Tables / Python functions / Excel formula, etc., such that the **shaded area under the curve is 0.95**.
    - The interval {x1,x2} is said to constitute the **95% Confidence Interval (95% CI)**.
    - The interpretation is that all values between x1 and x2 are **statistically not really different from the REFERENCE VALUE**.
    - Why? - Because, if the population parameter equals the REFERENCE VALUE, and If you were to take 100 random samples (si) from the population, 95 of those samples will result in the calculated value of the ***statistic*** lying between x1 and x2.



- Now we come back to our critical question, posed earlier, and ask it in the form of the following **Statistical Hypothesis: "Is a1 statistically different from ZERO"**
    - We want to check if the statistic **a1** is significantly different from ZERO
    - **Hence we choose the reference value to be ZERO**,
    - **Now, does the 95% CI {x1, x2} include the calculated value of the coefficient a1 (ie. 5.4522)**?
    - If YES, then the calculated value 5.4522 is **obtained by chance** and it is **NOT statistically different from ZERO** - hence our regression model is not valid.
    - If NO, then we can confidently say the following:
        - That the calculated value 5.4522 is SIGNIFICANTLY DIFFERENT from ZERO.
        - That, the value of 5.4522 is not resulting from chance because of the peculiarity of the specific sample s1
        - That 95 out of 100 random data samples would also have resulted in a value of ai closer to 5.4522.
    - All this is equivalent to checking whether the calculated value (eg. a1 = 5.4522) lies within the green region (the 95% region) around ZERO or, does it lie in the red region (the 5% region)?
        - If within the 95% (green) region: The calculated value is statistically not different from the reference value ZERO
        - If outside the 95% region (ie. within the remaining 5% region): The calculated value is statistically different from the reference value ZERO - and hence stated to be SIGNIFICANT, RELEVANT and VALID



- So, finally, **what is p-value and how should it be interpreted**?
    - We would like to check what is the position of **a1** (= 5.4522) with respect to the reference value (ZERO) of the established Sampling Distribution of **a**, and this we do by calculating the sum of the area(s) under the curve to the far sides of a1, and then comparing it to the reference value of 0.05 (ie. 5%).
    - **p-value** is defined as the **sum** of the areas under the curve beyond the far side of the calculated value, and the corresponding area on the opposite side - since a1 can take positive and negative values. For example, in the figure alongside, the red shaded area to the right lies beyond the calculated value of a1, and this area is 2.36E-38. Similar sized area is also considered to the left, and these two add to a total of 4.72E-38, which is the p-value value reported in the regression output. This belongs to the class of Hypothesis Tests known as the **two-tailed** tests because a1 can take both positive and negative values.
        - In some Hypothesis Tests, the region of interest is on only one side, and such tests are known as **one-tailed** tests.
    - As this p-value is much less than the threshold value of 0.05 (ie. 5%), we can conclude that 5.4522 is outside the 95% confidence interval for the reference value of ZERO, hence it STATISTICALLY SIGNIFICANT (ie. statistically different from ZERO) and we can rely on it.
    - If the p-value was greater than 0.05, it would necessarily imply that the calculated point 5.4522 is actually lying inside the 95% CI (the green zone), hence NOT STATISTICALLY SIGNIFICANT (ie. not different from ZERO).
    - As we will see later, in MLR, the p-values help us to **select** the most appropriate independent variables (**x**'s) to be included in the model.



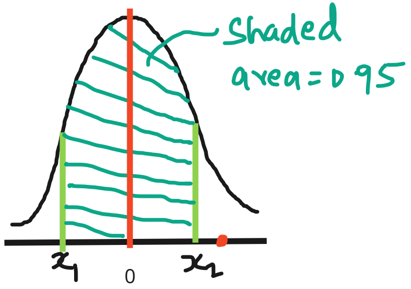
LR output also calculates x1 and x2 with the sampling distribution centered around a1 = 5.4522.
In the example:
x1 = 4.9398
x2 = 5.9647

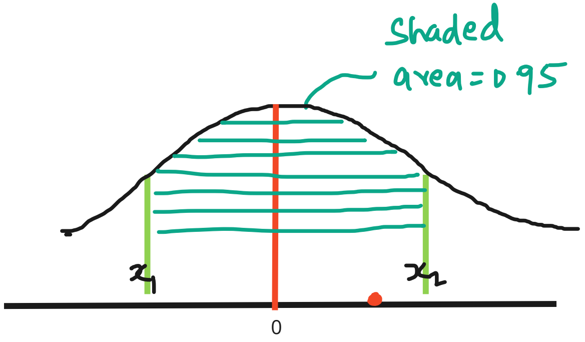
It is worth noting that ZERO is NOT a part of this interval.

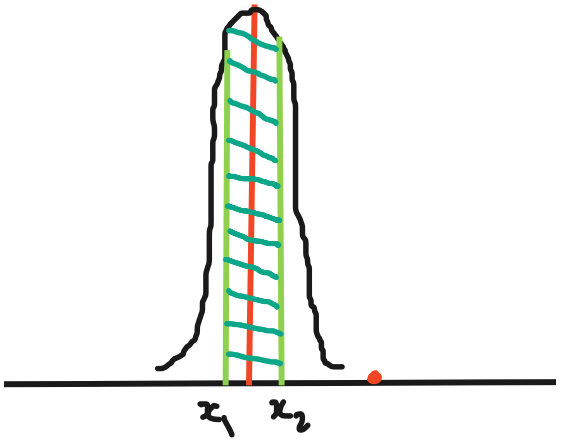
This also tells us that 95 out of 100 samples **si** will result in an **ai** value lying within this interval

The implications of sample size on the confidence interval:

- It should be noted that the formulae for standard errors have the sample size **n** in the denominator.
- Consequently, **if the sample size is reduced** the standard error value becomes larger and the sampling distribution curve gets stouter and wider as shown alongside, and the x1 and x2 values get pushed further away from ZERO
- The net effect is that values such as the red dot, which was SIGNIFICANT earlier now falls within the **not statistically different from ZERO zone**! The zone of uncertainty has widened, and it is more difficult to get a valid model.
- Conversely, if the sample size is increased, the standard error value becomes smaller and the sampling distribution curve gets slimmer and the x1, x2 values get pulled towards ZERO - the range of values considered *statistically equal to ZERO* becomes small.
    - This increases the chance that the red dot (calculated a1) will lie far outside the 95% CI for ZERO and hence it will considered as statistically significant. In effect, **uncertainty in the model reduces as the sample size increases**.







What about **F** and *SIgnificance F?*

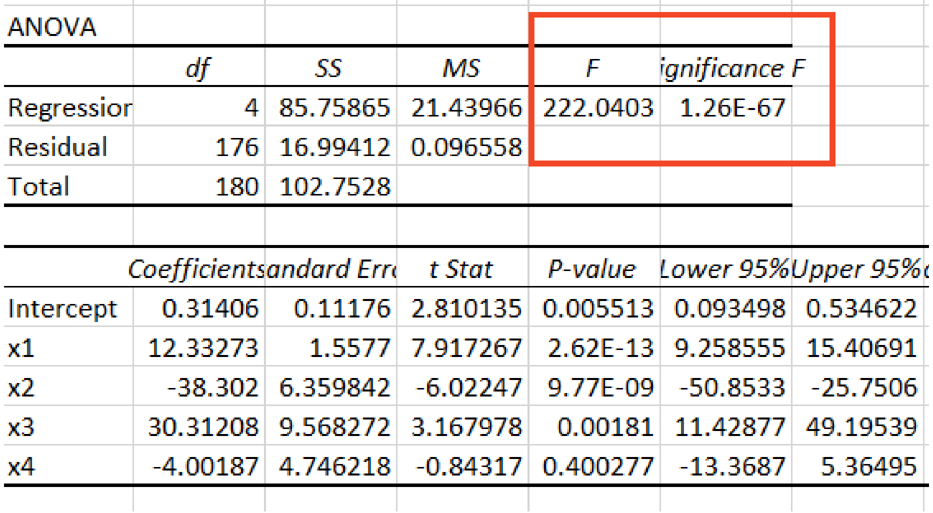
- These are relevant in case of Multiple Linear Regression (MLR) where **y = f(x1, x2, x3, x4, ...)**
- The **F-statistic** is defined as follows:

$$F = \frac{\frac{SSR}{n-k-1}}{\frac{SSE}{n-1}}$$

n is the number of observations,

k is the number of predictors (independent variables) in the model.

- As we can see, it is the ratio of **average variance explained by regression** and **average variance attributable to random errors (MSR / MSE)**
- The better the regression model, the larger will be the value of the **F-statistic** - that is, significantly more variance in the data will be explained by the regression model.
- **Significance F** is nothing but the p-value associated with the **F-statistic** Therefore, if **Significance F** is less than 0.05 (5%) we can rely on the calculated **F-statistic** and consider it **statistically significant** and use it to confidently make judgements about the overall quality of the Linear Regression model.
- ***We assess the LR model quality by taking into account the calculated coefficients (and their p-values) and the F-statistic (and it's p-value)***
- This is not the end !! There are a few more metrics, that we will encounter soon ...



ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	4	85.75865	21.43966	222.0403	1.26E-67
Residual	176	16.99412	0.096558		
Total	180	102.7528			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.31406	0.11176	2.810135	0.005513	0.093498	0.534622
x1	12.33273	1.5577	7.917267	2.62E-13	9.258555	15.40691
x2	-38.302	6.359842	-6.02247	9.77E-09	-50.8533	-25.7506
x3	30.31208	9.568272	3.167978	0.00181	11.42877	49.19539
x4	-4.00187	4.746218	-0.84317	0.400277	-13.3687	5.36495