



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA

**Universidad Politécnica de Valencia**

**Escuela Técnica Superior de Ingeniería Informática**

# STATISTICAL STUDY OF THE STACKOVERFLOW'S SURVEY 2022 FOR THE IBERIAN PENINSULA.



---

**Author: Rodríguez Díaz, Gabriel**

**Tutor: Zarzo Castelló, Manuel**

## **Abstract**

The Stack Overflow Survey of 2022 is a comprehensive dataset that provides valuable insights into the programming community. This statistics project aims to analyze and explore various aspects of the survey data to gain a deeper understanding of the trends, preferences, and demographics of software developers of the Iberian Peninsula.

The project begins by cleaning and preprocessing the survey data using scripts to ensure consistency and accuracy. Exploratory data analysis techniques are then applied to examine the distribution of variables such as blockchain opinion, salary, years working as developer... Descriptive statistics and visualizations are utilized to identify patterns, trends, and correlations within the dataset.

Furthermore, the project investigates the relationship between different variables, employing inferential statistical methods such as hypothesis testing and regression analysis. This allows for the examination of factors that influence salary levels. Additionally, the project explores potential differences and similarities among developers from Spain and Portugal.

Moreover, the project aims to address specific research questions. By conducting statistical analyses and interpreting the results, this project contributes to a comprehensive understanding of the Stack Overflow Survey of 2022, shedding light on the current state of the industry and providing valuable insights for developers, employers, and policy-makers alike.

Overall, this statistics project offers a detailed exploration of the Stack Overflow Survey of 2022, utilizing various statistical techniques to uncover significant patterns and relationships within the dataset or just to clarify doubts of possible relationships. The findings contribute to the collective knowledge of the programming community and offer actionable insights for individuals and organizations involved in the software development industry.

## Distributions in sampling - Inference about one population

**22. Assume that the variable YearsCode follows a normal distribution with average equal to the sample average and standard deviation equal to the sample value. If we take a random sample with 10 values and we compute its average, compute the confidence interval that would comprise the 95% of these values. (Solve theoretically)**

Assuming that the variable YearsCode follows a normal distribution

$N(\mu=14,2389 ; \sigma=8,80874)$  and taking 10 random values we have that  $N=10$ .

We know that  $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$  therefore, using the Z-Table, the interval will be

$$\bar{x} \in \mu \pm 2 \frac{\sigma}{\sqrt{n}} . \text{ Note that 2 would be around 1.96.}$$

That gives an interval of  $\bar{x} \in [8.7791, 19.6986]_{1-\alpha=95\%}$  |

**23. Assume that YearsCode follows a normal distribution. Calculate the percentile 52, ( $Z_{52}$ ).**

According to Statgraphics, the percentile 52,  $Z_{52}= 13.0$

**23.1 Solve the contrast between  $H_0: m = Z_{52}$  over  $H_1: m \neq Z_{52}$  considering as size of the sample the observations and significance level of 10%.**

$$\begin{cases} H_0: m = 13 \\ H_1: m \neq 13 \end{cases}$$

Considering  $n=494$  and  $\alpha=10\%$

Afterwards we obtain  $t_{n-1}$  from the table. Doing so, we get  $t_{494} = 0,999932$

Finally, we should apply the following formula:

$$\frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}} \sim t_{n-1} = \frac{14,2389-13}{\frac{8,80874}{\sqrt{494}}} \sim t_{493} = 3,1259$$

Since  $3,1259 > 0,999932$  we do not reject the null hypothesis for  $\alpha=10\%$

### 23.2 Solve the same hypothesis contrast with Statgraphics.

Using Statgraphics you have to follow the next steps:

1. *Describir > Datos numéricos > Análisis de una variable (Seleccionar YearsCode)> Marcar Test de hipótesis.*
2. Select **Prueba de hipótesis para YearsCode**
3. *Right click > Opciones de ventana*
4. Introduce 14,2389 in *Media/Mediana*
5. Introduce 10% in *Alpha*.

Once followed the previous steps, you just have to have a look at the results.

The following output will be shown:

Media Muestral = 14,2389  
 Mediana Muestral = 12,0  
 Desviación Estándar de la Muestra = 8,80874

Prueba t  
 Hipótesis Nula: media = 14,2389  
 Alternativa: no igual

Estadístico t = -0,0000847873  
 Valor-P = **0,999932**

No se rechaza la hipótesis nula para alfa = 0,1.

**24. Assume that YearsCode follows a normal distribution  $X \approx N(\mu; \sigma)$  with media and typical deviation equal to the sample's.**

**If a 10 sample from this population is randomly taken and the variance is obtained, calculate the confidence interval for the 95% of the values.**

Assuming that the variable YearsCode follows a normal distribution  $N(\mu=14,2389; \sigma=8,80874)$  and taking 10 random numbers, we have that  $N=10$ .

We know that  $(N-1) \frac{s^2}{\sigma^2}$  follows a Chi Square distribution with  $N-1$  degrees of freedom, for this case, 9.

We compute the interval that comprises the 95% of the values of a Chi Square distribution  $\chi^2_9$ . In this case  $[3,3251; 16,9189]$

Let's calculate the interval for which  $s^2$  belongs.  $\frac{9*s^2}{\sigma^2} \in [3,3251; 16,9189]$

$$s^2 \in \left[ \frac{\sigma^2_{3,3251}}{9}; \frac{\sigma^2_{16,9189}}{9} \right] = s^2 \in [28,6674; 145,867]$$

**25. Assume that YearsCode follows a normal distribution  $X \approx N(\mu; \sigma)$  with media and typical deviation equal to the sample's.**

**If a random sample of 12 data from the population is taken and the variance is obtained, what is the possibility of it being greater than  $3\sigma$ ?**

Assuming that the variable YearsCode follows a normal distribution  $N(\mu=14,2389; \sigma=8,80874)$  and taking 12 random numbers, we have that  $N=12$ .

We know that  $(N-1) \frac{s^2}{\sigma^2}$  follows a Chi Square distribution with  $N-1$  degrees of freedom, we can compute the probability we are asked to in the following way:

$$P\left(s^2 \frac{N-1}{\sigma^2} > 3\sigma \frac{N-1}{\sigma^2}\right) = P\left(X_{11}^2 > 3 \frac{11}{\sigma}\right) = P(X_{11}^2 > 3,746)$$

Finally, using Statgraphics, we can compute the probability mentioned above.

1. *Describir > Ajustes de distribuciones > Distribuciones de probabilidad > Chi cuadrado.*
2. G.L. (Grados de Libertad) with value  $N-1$ , in this case 11
3. Check in *Distribuciones acumuladas.*
4. *Right click > Opciones de ventana > Variable Aleatoria > 3,746*
5. The table will be computed and we take the value for *Área Cola Superior >*

Thanks to Statgraphics we have that  $P(X_{11}^2 > 3,746) = 0,976767$

The probability of the variance being greater than  $3\sigma$  is quite high.

**26. Using YearsCode, if two random samples of 14 values are taken, which is the probability of the variance of the second sample is three times the first's?**

Even if the exercise indicates to use YearsCode, in the end, this will be irrelevant, since we will not take  $\sigma^2$  into account. Now we will see why.

Taking two random samples  $s_1^2$  and  $s_2^2$  we will have to calculate  $\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \sim F_{n_1-1, n_2-1}$  but

$\sigma_1^2 = \sigma_2^2$ , and  $n_1 - 1 = n_2 - 1$  so, in the end we have  $\frac{s_1^2}{s_2^2} \sim F_{13}$

Now we use the following formula:

$$P(s^2 > 3s^2) = P\left(\frac{s_1^2}{s_2^2} > 3\right) = P(F_{13,13} > 3) = 0,029$$

As we mentioned before, the dataset used does not change the probability of the second sample's variance being 3 times the first's.

## 27. Obtain a Confidence Interval of the 99% for the average of YearsCode.

The average for YearsCode is  $\mu=14,2389$ , we can find the Confidence Interval of 99% for this media with Statgraphics following the next steps:

1. *Describir > Datos numéricos > Análisis de una variable > YearsCode > Check Intervalos de Confianza*
2. *Right Click > Opciones de ventana > Intervalo de Confianza = 99%*
3. Take the confidence interval for *media*.

Now we have got the confidence interval for the average, which is

$$\mu \in [13,214; 15,2637] \text{ or } \mu \pm 1,02483$$

### 27.1 What would happen in case YearsCode did not fit a normal distribution?

This would mean that the calculus made are not valid.

### 27.2 “If any value belonging to the confidence interval is taken and an hypothesis test is performed over the average, the conclusion will be always the same taking $\alpha=1\%$ ” Is it true? Why?

For any value m inside the range the conclusion is not reject. Therefore, this is true.

## 28. Using YearsCode, obtain with Statgraphics a confidence interval of 95% for the typical deviation of YearsCode. Calculate the Interval with a 99%

The typical deviation for YearsCode is  $\sigma=8,80874$ , we can find the Confidence Interval of 99% for this media with Statgraphics following the next steps:

1. *Describir > Datos numéricos > Análisis de una variable > YearsCode > Check Intervalos de Confianza*
2. *Right Click > Opciones de ventana > Intervalo de Confianza = 95% or 99%*
3. Take the confidence interval for *desviación típica*.

Now we have got the confidence interval for the typical deviation, which is

For 95%  $\sigma \in [8,29157; 9,39525]$

For 99%  $\sigma \in [8,13779; 9,59095]$

### 28.1 Which interpretation does it has?

In this case, for  $\alpha=5\%$  the interpretation is just how wide we want the range to be.

### 28.2 Which interval looks better? Which are the factors?

For this case it does not really matter. If we were following an efficiency or economic criteria, like it may be used in a factory, we should take a smaller range, that is, a bigger alpha.

We may note that in industry the most followed criteria tend to be

$$m \pm 3\sigma \rightarrow 99,73\%$$

29. Indicate in a table the variance of YearsCode and the amount of data for each one of the variants of Country.

Country	YearsCode ( $\sigma^2$ )	Observations (N)
Spain	78,7185	402
Portugal	73,1671	92

TABLE 14: Variance of YearsCode and Observations for the subset Country.

### 29.1 Can we estate that the differences are statistically significant?

Considering  $\alpha=0.05$  for these computations.

Let's define the null hypothesis:  $H_0: \sigma_{Spain}^2 = \sigma_{Portugal}^2$

We have to compute the ratio  $\frac{\sigma_{Spain}^2}{\sigma_{Portugal}^2}$

To do so, we have to compute the confidence interval and evaluate the ratio against the confidence interval  $\frac{\sigma_{Spain}^2}{\sigma_{Portugal}^2} = 1.075 \sim F_{N_{Spain}-1, N_{Portugal}-1}$

Now, using the table, we get the following result:

$$1.075 \sim F_{401, 91} = 1.419$$

Since the value  $1.075 \leq 1.419$  we cannot reject the null hypothesis, therefore there are not significant differences.



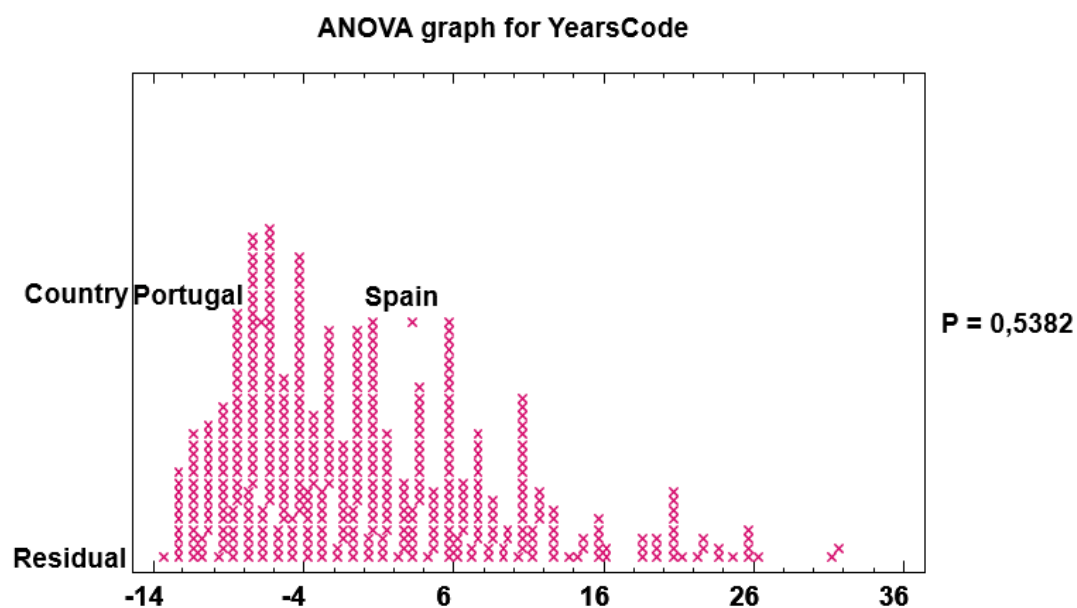
## ANOVA – Analysis of Variance

### 30. Make and ANOVA to study the effect of Country on the variable YearsCode.

Firstable, it is important to remark that a 0,05 significance level was chosen. The main reasons are the following:

- The study does not require a high level of confidence. It is just a simple study to know the relationship between variables. This information is not relevant for any other purpose that may require a high precision.
- The sample size also matters. Using a huge sample size is more difficult to determine smaller effects. That is why a smaller significance level may be used. For this case, it is not a problem since the sample is small.
- It provides a better balance between making type I and type II errors.  $\alpha=0.01$  increases the possibility of accepting the null hypothesis when it is false.

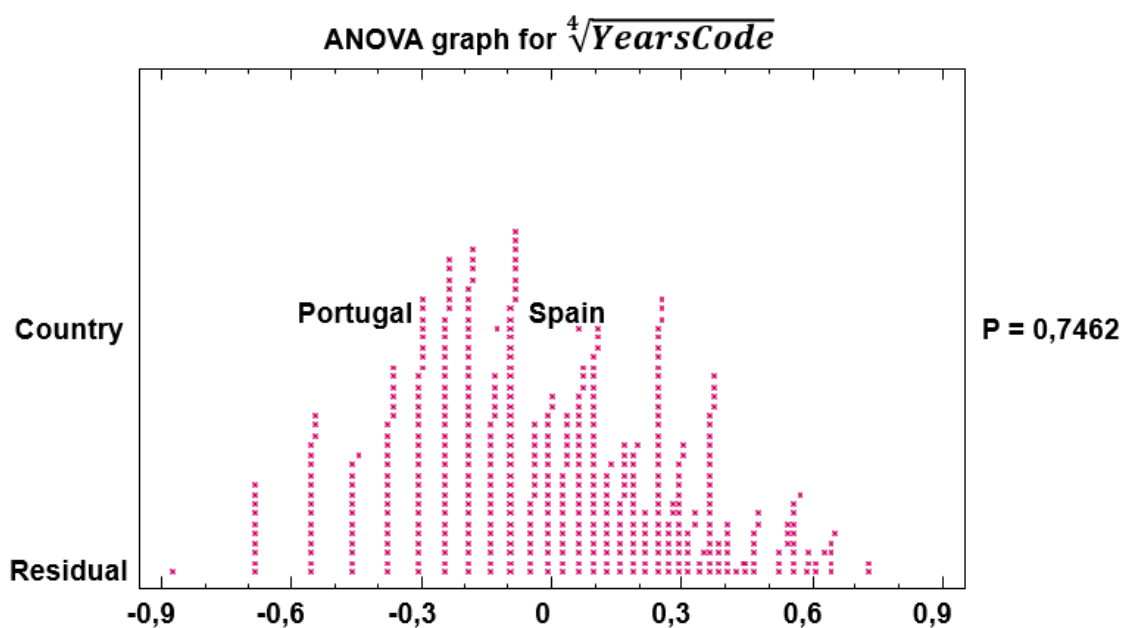
It was necessary to transform the data, because it was positive skewed.



PICTURE 31: Anova graph for YearsCode

As we can see, this graph is positively skewed, so some transformation will be performed.

To assess skewness, we can look at the Standarized Skewness values. Both Portugal and Spain have positive Std. Skewness values (5.359 and 7.56, respectively).



PICTURE 32: Anova graph for transformed YearsCode

For this case the total Skewness Coefficient is -0,405 which makes it the best option.

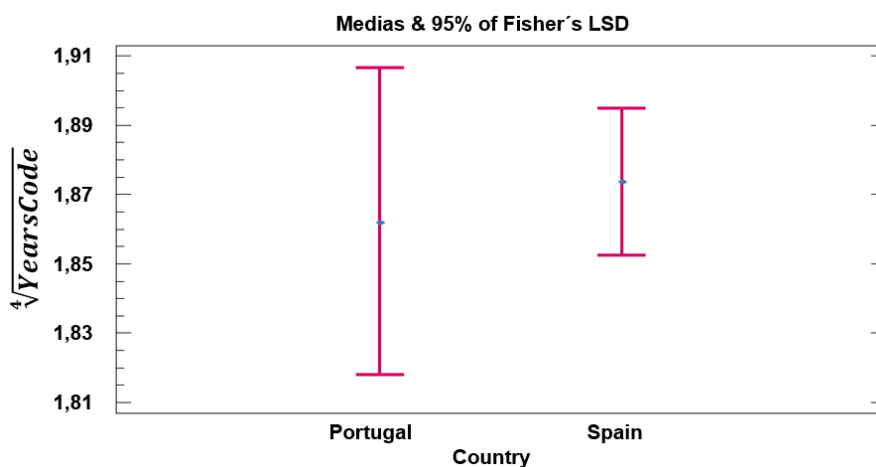
Now, we can have a look at the ANOVA table of the model.

Tabla ANOVA para  $\text{YearsCode}^{0.25}$  por Country

Source	Square Sum	Gl	Cuadrado Medio	Razón-F	Valor-P
Entre grupos	0,0098051	1	0,0098051	0,10	0,7462
Intra grupos	46,003	492	0,0935021		
Total (Corr.)	46,0128	493			

TABLE 14: ANOVA table for  $\sqrt[4]{\text{YearsCode}}$

It is also important to have a look at the LSD intervals:



PICTURE 33: Medias and Fisher's LSD Intervals for YearsCode transformed for Country

As we can see the intervals do not have the same amplitude. This is because these intervals are the LSD (Least Significant Differences) intervals, which are computed following this formula:

$\bar{x}_i \pm \frac{\sqrt{2}}{2} t_{d.f. \text{ res.}}^{\frac{\alpha}{2}} \sqrt{\frac{MS_{resid}}{K}}$  being  $\bar{x}_i$  the sample average corresponding to the  $i$  variant of the factor and  $K$ , the total number of data used to compute  $\bar{x}_i$ . Since this formula depends on both the average of the different variants and the data used, it is obvious that the intervals will be different for each variant.

Now, we may ask ourselves if the conclusions extracted from the graph are coherent with the ANOVA table (table 14).


Yes, they are. From the ANOVA table we know that there may be statistically significant differences between the averages because the P-Value is higher than  $\alpha=0,05$ . By looking at the LSD intervals (Picture 33), we can clearly see that the variants overlap. If they overlap it means that there are not statistically significant differences between them, but if they do not, then there are significant differences.

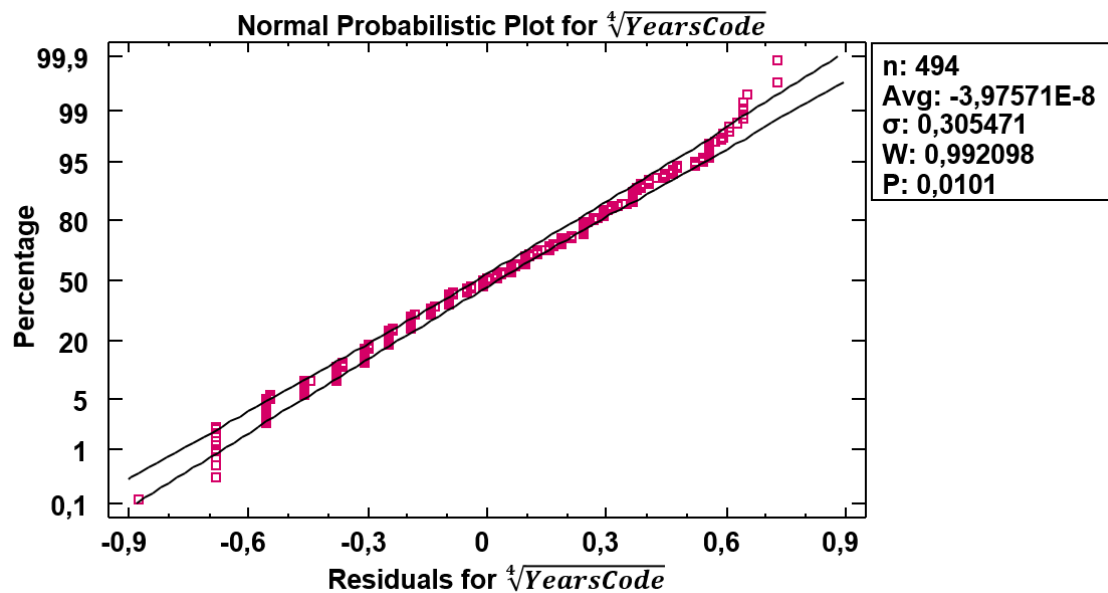
In summary, we can extract the following conclusions from this ANOVA:

As said before, having an  $\alpha=0.05$  and a P-Value higher than  $\alpha$ , means that there may be statistically significant differences between the mean of the variants. But we can not just look at the table. With the table (Table 14) it is not possible to know which variants means differ from the rest. That is why we have to take a look at the LSD intervals.

In the graphic (Picture 33), we can clearly see that the two intervals overlap, meaning that there is not statistically significant differences between the means.

Finally, we will plot a Normal Probabilistic Plot of the residual and discuss it.

To do so, in the same ANOVA Statfolio we are using, we have to click in this button  and check Residuos option to save the residuals with the name we desire. Afterwards, we go to *Graficar>Gráficos exploratorios>Papel probabilístico normal>RESIDUALS YearsCode*



PICTURE 34: Normal Probabilistic Plot for  $\sqrt[4]{\text{YearsCode}}$  Residuals

At first sight we might think that there are some outliers to be removed, nevertheless, having a look at the statistical summary made by Statgraphics, we have more information to take into account.

Standard Skewness Coefficient	-0,435872
Standard Kurtosis	-1,98759

TABLE 15: Statistical Summary for RESIDUALS  $\sqrt[4]{\text{YearsCode}}$

Also, the following output is shown:

De particular interés aquí son el sesgo estandarizado y la curtosis estandarizada, las cuales pueden utilizarse para determinar si la muestra proviene de una distribución normal. Valores de estos estadísticos fuera del rango de -2 a +2 indican desviaciones significativas de la normalidad, lo que tendería a invalidar cualquier prueba estadística con referencia a la desviación estándar. En este caso, el valor del sesgo estandarizado se encuentra dentro del rango esperado para datos provenientes una distribución normal. El valor de curtosis estandarizada se encuentra dentro del rango esperado para datos provenientes de una distribución normal

Therefore, we may state that, even if the Kurtosis is at the limit of being a significative deviation, we may consider this Normal Probabilistic Plot as valid and without significant outliers to be removed.

### 31. Incorporate to the previous model the factor Blockchain and the double interaction.

Source		Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Main Effects	A: Country	8,09928	1	8,09928	0,10	0,7481
	B: Blockchain	151,593	5	30,3185	0,39	0,8580
Interaction : AB		180,528	5	36,1056	0,46	0,8057
Residuals		37797,5	482	78,4181		
Total (Corrected)		38253,8	493			

TABLE 16: Anova Table for YearsCode by Country and Blockchain with double interaction

Source		Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Main Effects	A: Country	0,000800557	1	0,000800557	0,01	0,9266
	B: Blockchain	0,261471	5	0,0522942	0,56	0,7341
Interaction : AB		0,230103	5	0,0460207	0,49	0,7846
Residuals		45,3677	482	0,0941238		
Total (Corrected)		46,0128	493			

TABLE 17: Anova Table for  $\sqrt[4]{\text{YearsCode}}$  by Country and Blockchain with double interaction

To check if any of the factors are non-significant we need to compute  $F_{df. \text{factor}, df. \text{residual}}$  for each of the factors and check if it is bigger than the F-Ratio obtained in the table. For all of the computations we will use the same  $\alpha$  as in the previous exercise (0,05).

For the factor Country:

- $F_{1,482} = 3,86083852 > F\text{-Ratio} = 0,01$
- Country is not significant at 0.05 significance level.
- Reject null hypothesis.

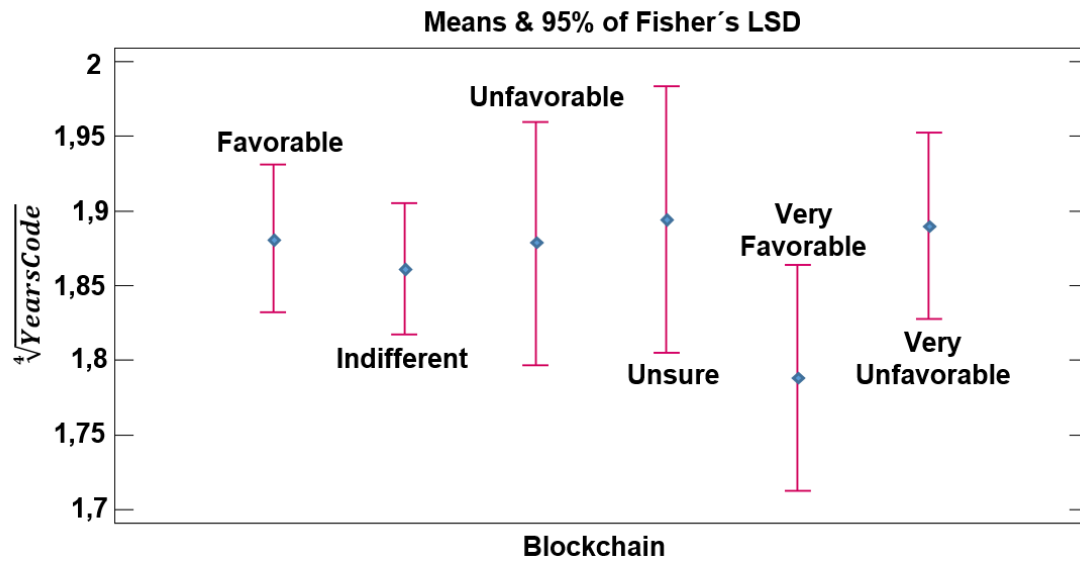
For the factor Blockchain

- $F_{5,482} = 2,232597359 > F\text{-Ratio} = 0,56$
- Blockchain is not significant at 0.05 significance level.
- Reject null hypothesis.

For the Interaction:

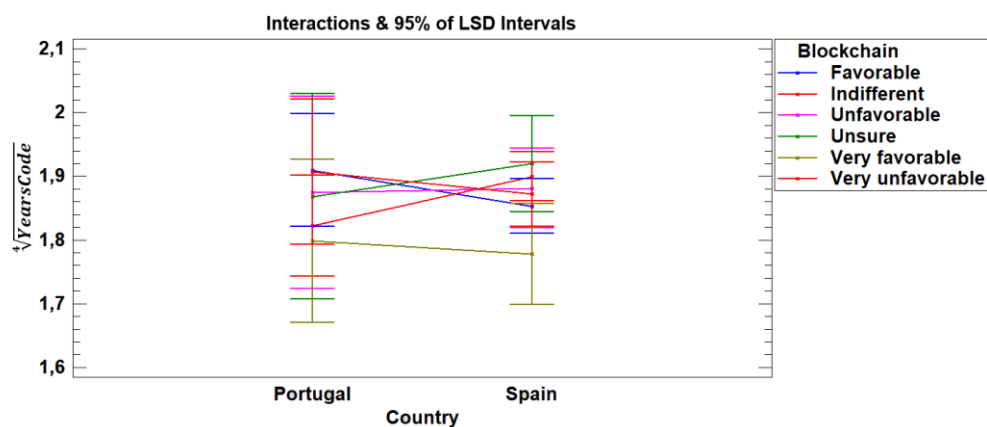
- $F_{5,482} = 2,232597359 > F\text{-Ratio} = 0,49$
- The interaction is not significant for 0.05 significance level.
- Reject null hypothesis.

Therefore, based on the given  $\alpha=0.05$ , the interpretations is that the three factors are not significant.



PICTURE 35: Medias and Fisher's LSD Intervals for  $\sqrt[4]{\text{YearsCode}}$  for Blockchain.

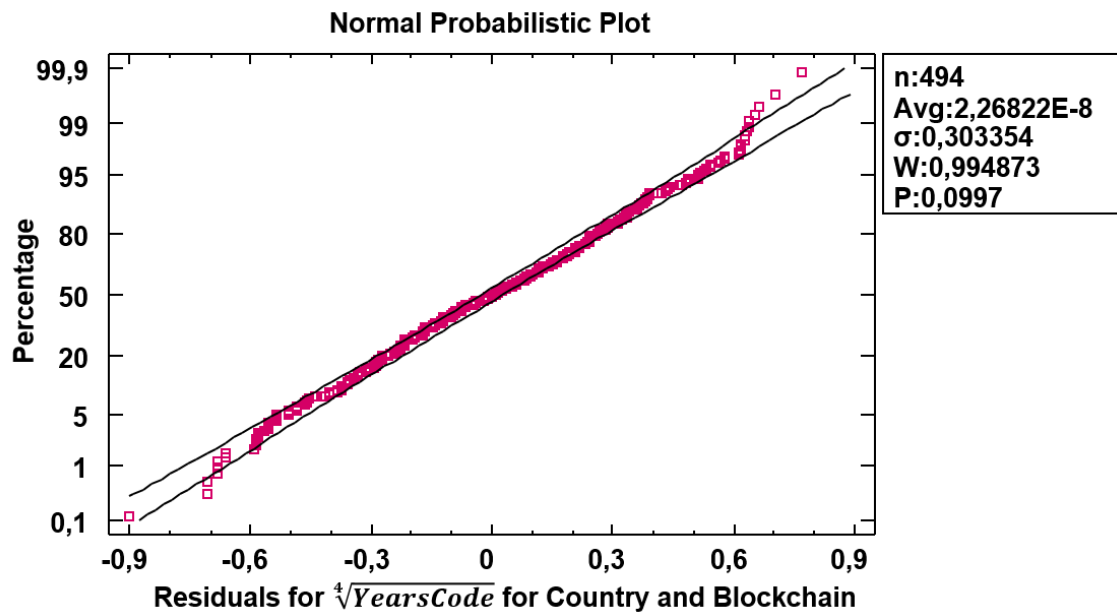
By looking at the LSD intervals (Picture 35), we can clearly see that the variants overlap. If they overlap it means that there are not statistically significant differences between them. So, the results coincide with the Table's (Table 17).



PICTURE 36: Interactions and LSD Intervals for  $\sqrt[4]{\text{YearsCode}}$  for Country and Blockchain.

In this graphic we can clearly see that the interactions are not significant since the lines do not tend to be parallel. We can also determine which means are equal since the points are very close together in some of them.

For instance, with Unfavorable and Unsure, we have a very similar mean since the points are in a very short distance from each other. Also for Very Unfavorable and Favorable the mean is practically the same since the points are even closer.



PICTURE 37: Normal Probabilistic Plot for  $\sqrt[4]{\text{YearsCode}}$  Residuals for Country and Blockchain

For this case, the analysis is quite similar to the previous done with Picture 34.

At first sight there are some possible outliers, but we have to take into account the following data:

<b>Standard Skewness Coefficient</b>	<b>-0,322556</b>
<b>Standard Kurtosis</b>	<b>-1,8863</b>

TABLE 18: Statistical Summary for  $\sqrt[4]{\text{YearsCode}}$  Residuals for Country and Blockchain

As we can see, the Kurtosis is even lower (Than table's 15) in absolute value. Then we may consider this Normal Probabilistic Plot as valid and without significant outliers to be removed.

### 32. Make and ANOVA to study the effect of the factors Country and Blockchain over YearsCode. Work with the original variable.

Again, the reasoning for choosing a significance level is the same as before. In summary the reasons would be:

- The study does not require a high level of confidence.
- The sample size is small.
- It provides a better balance between making type I and type II errors.

So, even if the significance level does not matter a lot for this case, the one chosen is 0.05

Source		Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Main Effects	A: Country	20,6143	1	20,6143	0,26	0,6074
	B: Blockchain	246,268	5	49,2536	0,63	0,6757
Residuals		37978,1	487	77,9837		
Total (Corrected)		38253,8	493			

TABLE 19: Anova Table for YearsCode by Country and Blockchain

To check if any of the factors are non-significant, we need to compute  $F_{df. factor, df. residual}$  of the factors and check if it is higher than the F-Ratio of the table. Let's check for  $\alpha=0.05$

For the factor Country:

- $F_{1,482} = 3,860638966 > F\text{-Ratio} = 0,10$
- Country is not significant at 0.05 significance level.
- Reject null hypothesis.

For the factor Blockchain

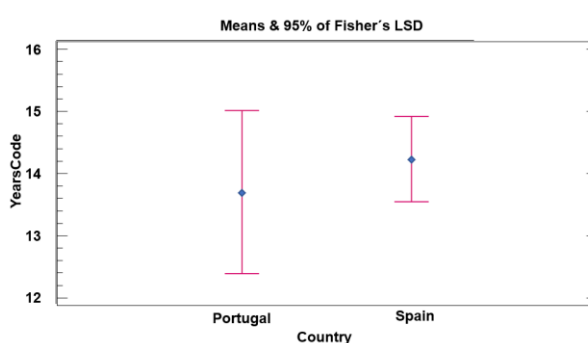
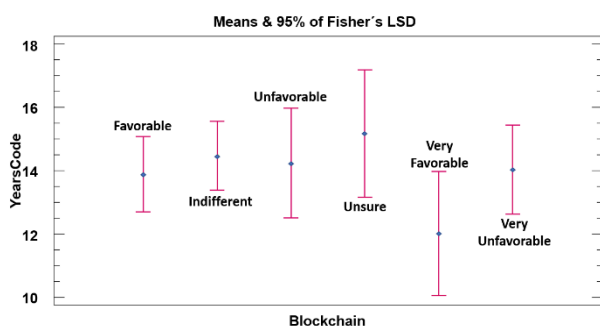
- $F_{5,482} = 2,232415603 > F\text{-Ratio} = 0,63$
- Blockchain is not significant at 0.05 significance level.
- Reject null hypothesis.

As we can see, they are not significant, now we may try with  $\alpha=0.01$

For Country  $F_{1,482} = 6,687770953$  it is still not significant.

For Blockchain  $F_{5,482} = 3,0547685$  not significant as well.





PICTURE 38: Medias and Fisher's LSD Intervals for YearsCode for Blockchain. PICTURE 39: Medias and Fisher's LSD Intervals for YearsCode for Country.

The conclusions obtained from the analysis are the following:

As we can clearly see, in the **Table 19** both P-Values are higher than  $\alpha$ , which means that there is no statistical evidence to support the presence of significant differences among the means of the groups being compared. Moreover, in both LSD graphs, the variants overlap. Which is another evidence.

In the case of Country, it is also remarkable that the variants with less years of code on average is the Spanish. This may be because the because of the specialized usage of Internet in Portugal, which is higher than in Spain. As we can see in the table below by having a look at the servers per inhabitant.

	Spain Total	Spain per 1000 inh.	Portugal Total	Portugal per 1000 inh.
Internet servers	4.228.000	89	3.748.000	363
Internet users	44.522.214	939	8,498.528	823

TABLE 20: Usage of the Internet for Spain and Portugal. Data from WorldData.info

On the other hand, for Blockchain, the variant with less years of code on average is Very Favorable and the one with most is Unsure. This may happen because the people with more ages of code tend to forget about the trends and the new technologies as they get older.

As it was said before, the values are not significant for  $\alpha=0.01$  nor  $\alpha=0.05$ . For this results it is possible to take the conclusion that the Null Hypothesis is true. Meaning that there are genuinely no significant differences among the groups being compared. In this case, obtaining non-significant results is expected.

In the end, there are a lot of variability in the ages of the people, different ideologies and life experiences. That could lead someone to reject blockchain technology, or also,

on the other hand, for Country, it is hard to analyze the impact of the years coding without taking into account the ages of the population.

### 33. Incorporate the double interaction to the previous model.

Source		Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Main Effects	A: Country	8,09928	1	8,09928	0,10	0,7481
	B: Blockchain	151,593	5	30,3185	0,39	0,8580
Interaction: AB		180,528	5	36,1056	0,46	0,8057
Residuals		37797,5	482	78,4181		
Total (Corrected)		38253,8	493			

TABLE 21: Anova Table for YearsCode by Country and Blockchain with double interaction.

To check if any of the factors are non-significant, we need to compute  $F_{df. factor, df. residual}$  of the factors and check if it is higher than the F-Ratio of the table. Let's check for  $\alpha=0.05$

For the factor Country:

- $F_{1,482} = 3,86083852 > F\text{-Ratio} = 0,10$
- Country is not significant at 0.05 significance level.
- Reject null hypothesis.

For the factor Blockchain

- $F_{5,482} = 2,232597359 > F\text{-Ratio} = 0,39$
- Blockchain is not significant at 0.05 significance level.
- Reject null hypothesis.

For the factor AB

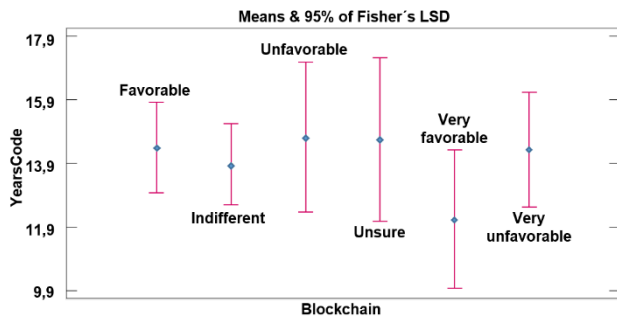
- $F_{5,482} = 2,232597359 > F\text{-Ratio} = 0,39$
- AB is not significant at 0.05 significance level.
- Reject null hypothesis.

As we can see, they are not significant, now we may try with  $\alpha=0.01$

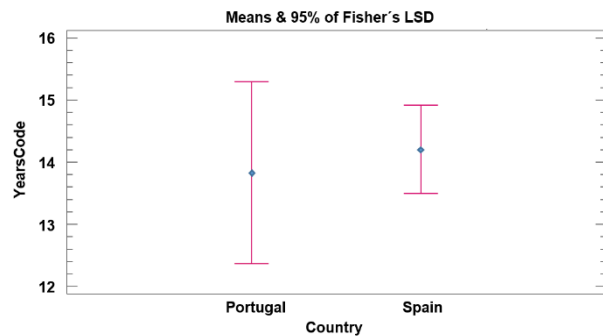
For Country  $F_{1,482} = 6,687770953$  it is still not significant.

For Blockchain and AB  $F_{5,482} = 3,0547685$  not significant as well.

As we can see, the data may vary a little, nevertheless, the conclusions extracted are the same. Even counting the factor AB it is not relevant. So, in the end. The result does not change if we add the interaction.



PICTURE 40: Medias and Fisher's LSD Intervals for YearsCode for Blockchain. With double interaction.



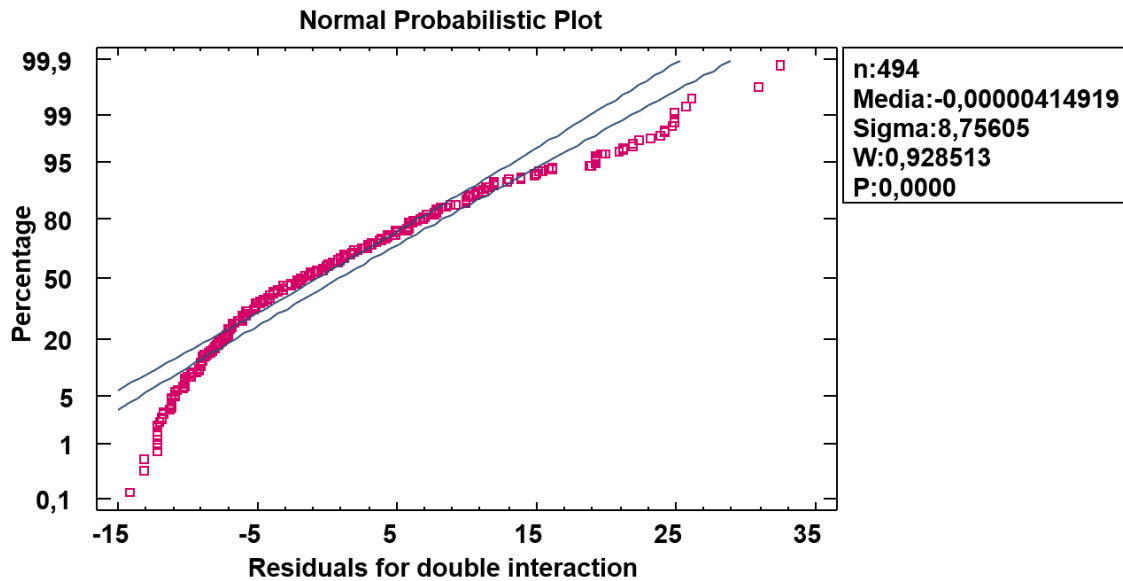
PICTURE 41: Medias and Fisher's LSD Intervals for YearsCode for Country. With double interaction.

There are cases where studying the interaction between factors may not make sense or may not be meaningful. In this case, there is strong prior knowledge suggesting that the factors do not interact or that the interaction is not of interest.

Nevertheless, personally I considered it interesting to study it, because the view of Blockchain is more positive in Portugal than in Spain. So, both factors could be conceptually related.

It is possible to observe little changes by staring at both LSD graphs pairs (Pictures 38,39,40,41). The most remarkable is the increasing in the mean for Unsure in Blockchain factor. But this does not give a lot information per se.

Even if for this case, no values are significant, we are taking the residuals and representing them in a Normal Probabilistic Plot.



PICTURE 42: Normal Probabilistic Plot for the residuals in double interaction.

If the plot of residuals deviates from a straight line or exhibits an asymmetric pattern, it suggests that the residuals do not follow a normal distribution. The shape and pattern of the plot can provide insights into the specific departure from normality.

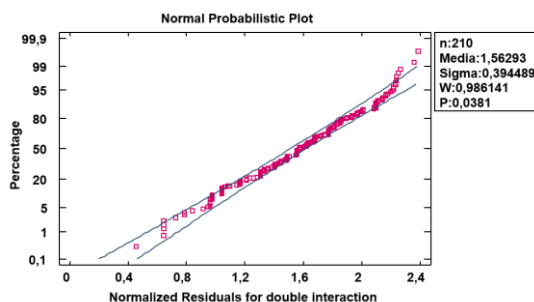
The plot shows a skewed pattern, with points bending away from and towards the straight line. It suggests that the residuals are not symmetrically distributed. Positive skewness indicates a longer tail to the right, which is this case. While negative skewness would indicate a longer tail to the left.

<b>Standard Skewness Coefficient</b>	<b>9,05275</b>
<b>Standard Kurtosis</b>	<b>3,38679</b>

TABLE 22: Statistical Summary for Residuals for Country and Blockchain with double interaction

For this case it could be possible to try to normalize it. It could indicate us that prior transformations would be convenient in the model.

For instance, by applying a fourth root transformation, the distribution can get more or less normalized.



PICTURE 43: Normal Probabilistic Plot for the transformed residuals in double interaction.

### 34. Make an ANOVA to study the effect of the factors of Country and Blockchain and its double interaction over YearsCodePro.

Source		Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Main Effects	A: Country	57,8382	1	57,8382	1,08	0,2988
	B: Blockchain	189,967	5	37,9934	0,71	0,6157
Interaction: AB		290,277	5	58,0554	1,09	0,3673
Residuals		25074	469	53,4628		
Total (Corrected)		25647,8	480			

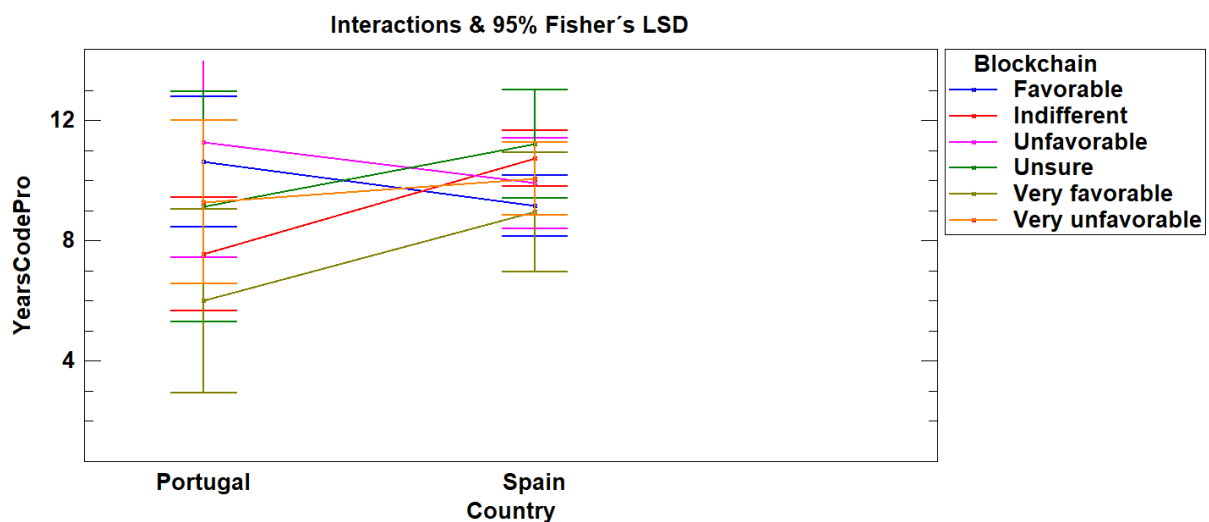
TABLE 23: Anova Table for YearsCodePro by Country and Blockchain with double interaction.

YearsCodePro is very similar to YearsCode in many ways. That is why the significance level chosen will be the same. Also, the reasoning is the same as the previous.

Now we are interpreting if the effect of the interaction AB is significant:

$$F_{5,469} = 2,233125743 > F\text{-Ratio} = 0,3673$$

This indicates that the interaction is not statistically significant. Let's review the LSD graphic for the two factors.



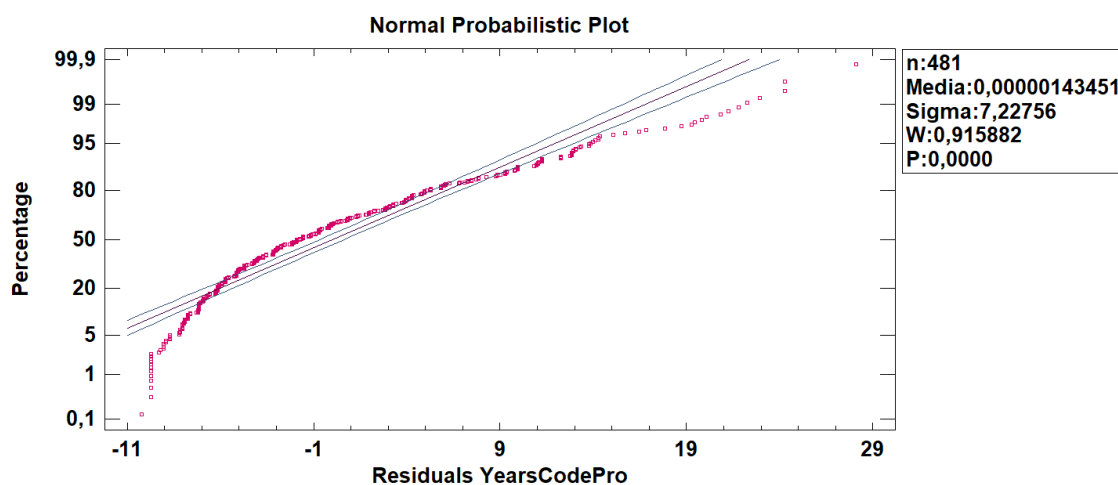
PICTURE 43: Interactions and LSD Intervals for YearsCodePro for Country and Blockchain.

In this graphic we can see that some of the interactions may be significant since the lines are more or less parallel. For the case of Unfavorable and Favorable, which is really interesting. Also, it is remarkable the case of indifferent and Very Favorable.

We can also determine which means are equal since the points are very close together in some of them.

For instance, with Very Unfavorable and Unsure, we have a very similar mean since the points are in a very short distance from each other.

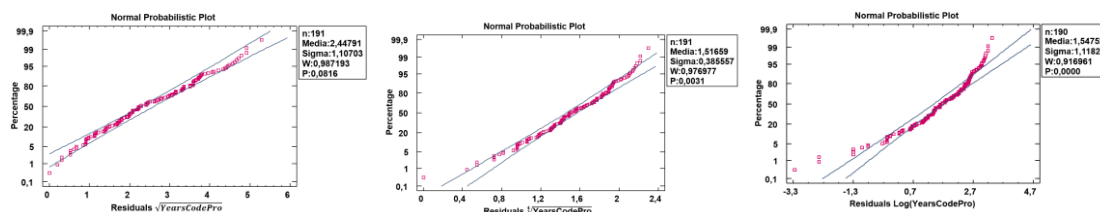
### 35. Study of the residuals of the previous model.



PICTURE 44: Normal Probabilistic Plot for YearsCodePro Residuals for Country and Blockchain.

The Normal Probabilistic Plot indicates us the distribution of the residuals. Residuals are the differences between the observed and predicted values of the response variable, and they reflect how well the model fits the data.

As we can see, the residuals do not follow a normal distribution, so we need to study which transformation is the best to normalize the residuals of the model, and then check if there are any outliers.



PICTURE 45,46,47: Normal Probabilistic Plots of transformed YearsCodePro Residuals for Country and Blockchain.

At first sight we can discard the logarithmic transformation. On the other hand, to choose between square and fourth root, we will choose in base of the Standard Skewness Coefficient and Kurtosis. Which is quite lower (in absolute value) in the case of the square root.

	Square Root	Fourth Root
Standard Skewness Coefficient	1,20488	-3,28397
Standard Kurtosis	-1,38675	1,95772

TABLE 24: Coefficients comparison between transformations

There are no outliers that should be erased from the study because all the points follow the same curve more or less and there are no values that clearly differ from the rest. Also, the Standard Skewness Coefficient and the Standard Kurtosis are relatively low (less than 2 points) Therefore, there are no abnormal values in the original variable YearsCodePro that should be erased because they are affecting the study negatively.

In summary, the conclusions that we got previously are valid and we do not need to change any of them.

## Linear Regression

**36. Obtain the matrix of variances-covariances for the variables YearsCode, YearsCodePro, Salary and AvgAge. What useful information does this matrix give? Why is it symmetric?**

	YearsCode	N	YearsCodePro	N	Salary	N	AvgAge	N
YearsCode	77,5939	494	58,4935	481	83211,1	494	56,1393	494
YearsCodePro	58,4935	481	53,4329	481	70110,7	481	47,8954	481
Salary	83211,1	494	70110,7	481	$8,32 \cdot 10^8$	495	71971,6	495
AvgAge	56,1393	494	47,8954	481	71971,6	495	71,6008	495

TABLE 25: Matrix of Variances-Covariances in table form.

$$\begin{pmatrix} 77,5939 & 58,4935 & 83211,1 & 56,1393 \\ 58,4935 & 53,4329 & 70110,7 & 47,8954 \\ 83211,1 & 70110,7 & 8,32 \cdot 10^8 & 71971,6 \\ 56,1393 & 47,8954 & 71971,6 & 71,6008 \end{pmatrix}$$

FIGURE 1: Matrix of Variances-Covariances.

This matrix gives us the information about the covariances of the variables with respect to one another, but it also provides us with the variances for each variable. This is because in the main diagonal of the matrix, what appears are the variances of each variable since the row and column coincide. This matrix is symmetric because we are representing the variables on the same order in the columns and rows and the covariance of X1 with respect to X2 is the same as the covariance of X2 with respect to X1, then the matrix will be symmetric because when we compute the covariance of two variables, it will not matter which one is in the row and which one is in the column, the result will be the same.



**37. Obtain the matrix of correlation of these variables. In the case of positive asymmetry, normalize the variables.**

	YearsCode	YearsCodePro	Salary	AvgAge
YearsCode		0,9151	0,3274	0,7527
		(481)	(494)	(494)
		0	0	0
YearsCodePro	0,9151		0,3312	0,7812
	(481)		(481)	(481)
	0		0	0
Salary	0,3274	0,3312		0,2948
	(494)	(481)		(495)
	0	0		0
AvgAge	0,7527	0,7812	0,2948	
	(494)	(481)	(495)	
	0	0	0	

TABLE 26: Matrix of correlation.

In order to know which transformations are the best, let's use the same criteria as before and have a look at the absolute values of the coefficient of Skewness and Kurtosis.

Remark that YearsCodePro has been analysed before and we already know the best transformation.

	Default	Square Root	Fourth Root	Logarithm
Standard Skewness Coefficient	9,04454	2,96612	-0,405685	-4,20632
Standard Kurtosis	3,04172	-1,83545	-2,02611	0,00446642

TABLE 27: Coefficients comparison between transformations of YearsCode.

	Default	Square Root	Fourth Root	Logarithm
Standard Skewness Coefficient	16,6149	1,60498	-21,2784	-35,4931
Standard Kurtosis	23,8612	11,609	56,7082	176,652

TABLE 28: Coefficients comparison between transformations of Salary.

	Default	Square Root	Fourth Root	Logarithm
Standard Skewness Coefficient	4,923	2,0514	0,642408	-0,751562
Standard Kurtosis	0,69144	-1,02276	-1,41527	-1,5271

TABLE 29: Coefficients comparison between transformations of AvgAge.

The conclusions are the following:

- YearsCode has to use a fourth root transformation.
- Salary has a really high skewness even with transformations. It would not mind, but a square root transformation will be applied.
- For AvgAge the best option is a fourth root.

	$\sqrt[4]{\text{YearsCode}}$	$\sqrt{\text{YearsCodePro}}$	$\sqrt{\text{Salary}}$	$\sqrt[4]{\text{AvgAge}}$
$\sqrt[4]{\text{YearsCode}}$		0,9009	0,3377	0,6991
		(481)	(494)	(494)
		0	0	0
$\sqrt{\text{YearsCodePro}}$	0,9009		0,3447	0,7606
	(481)		(481)	(481)
	0		0	0
$\sqrt{\text{Salary}}$	0,3377	0,3447		0,3013
	(494)	(481)		(495)
	0	0		0
$\sqrt[4]{\text{AvgAge}}$	0,6991	0,7606	0,3013	
	(494)	(481)	(495)	
	0	0	0	

TABLE 30: Matrix of correlation with transformations.

This matrix shows partial correlation coefficients between each pair of variables. For each pair, there are three values, the first one being the correlation coefficient, the second one the number of data and the third one the P-Value that is useful to know if the correlation is statistically significant.

In this case, it is used a confidence level of 95% which means that if the P-Value for one pair is less than 0,05 then it means that the correlation is significant.

As we can see, all the values here are correlated. This is because we are comparing three aging factors which are obviously correlated and another one which is the salary. It is clearly correlated since, in general, the higher the years coding and the age, the higher the experience, therefore, the higher the salary.

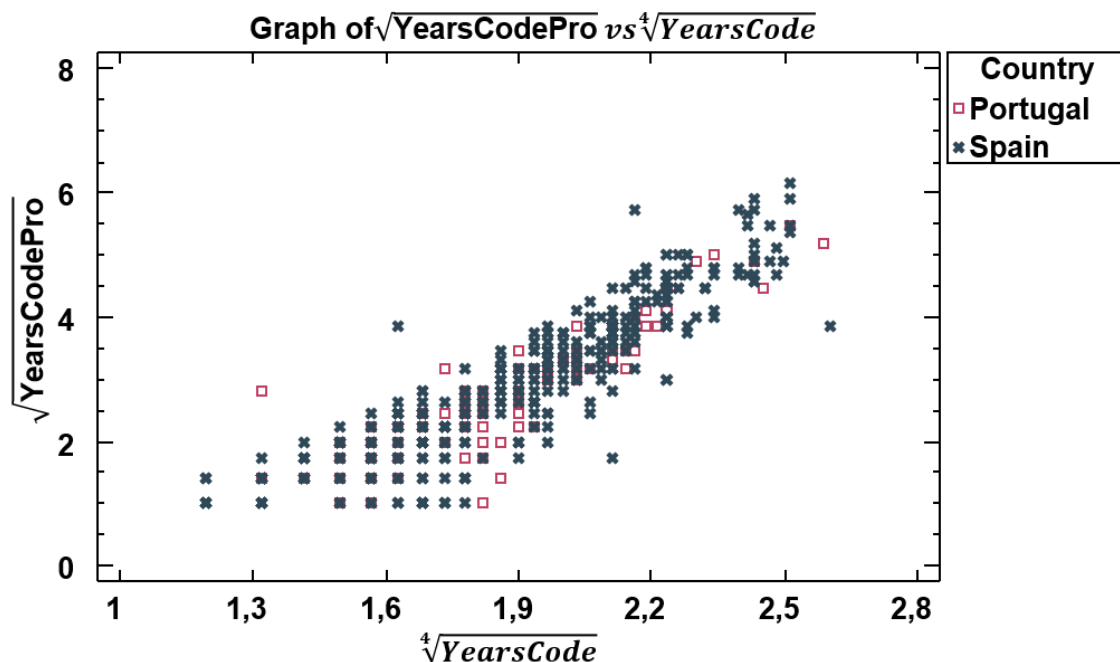
Taking into account the values, we see that the P-Value is zero for every variable, which is higher than the 0.05 threshold. On the other hand, we can see that the correlation coefficient is close to one in general for the age related variables, specially for YearsCode and YearsCodePro.

As it was stated previously in this assignment, the people tend to start coding as pro around after 2 years of starting.

The main diagonal is empty because it would not give us any useful information. This is because the values of the correlation coefficients of the main diagonal are always one, since the correlation coefficient of one variable with itself will always be one. This is why Statgraphics does not show any values on the main diagonal of this matrix.

**38. From the previous matrix, identify the couple of variables with a greater degree of correlations and plot a dispersion graphic between both.**

For this case  $\sqrt[4]{\text{YearsCode}}$  and  $\sqrt{\text{YearsCodePro}}$  are the ones with higher correlation with 0,9009. A little smaller than the limit of 0.95 established.

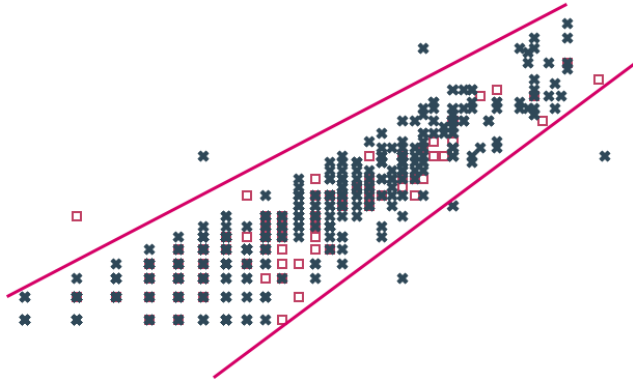


PICTURE 48: Dispersion graph of  $\sqrt[4]{\text{YearsCode}}$   $\sqrt{\text{YearsCodePro}}$ .

From the graphic we can see that both variables are correlated since, when the YearsCode is higher, the number of YearsCodePro also increases. This makes a lot of sense as it was explained before. People spend some time learning to code before start working on it. We can also see that most of the people with higher score years coding are the Spanish. This is obvious since the population of Spain is more than four times the Portugal's.

These variables have a positively linear relation, since we can fit a straight line to the plot, and this line would have positive slope. Also, we can see that both countries have a moderate correlation.

The points are really close from each other in general and the ones which are not so close, are not that far to consider it weak. Also it breaks the homoscedasticity hypothesis, as we can see in the image below.



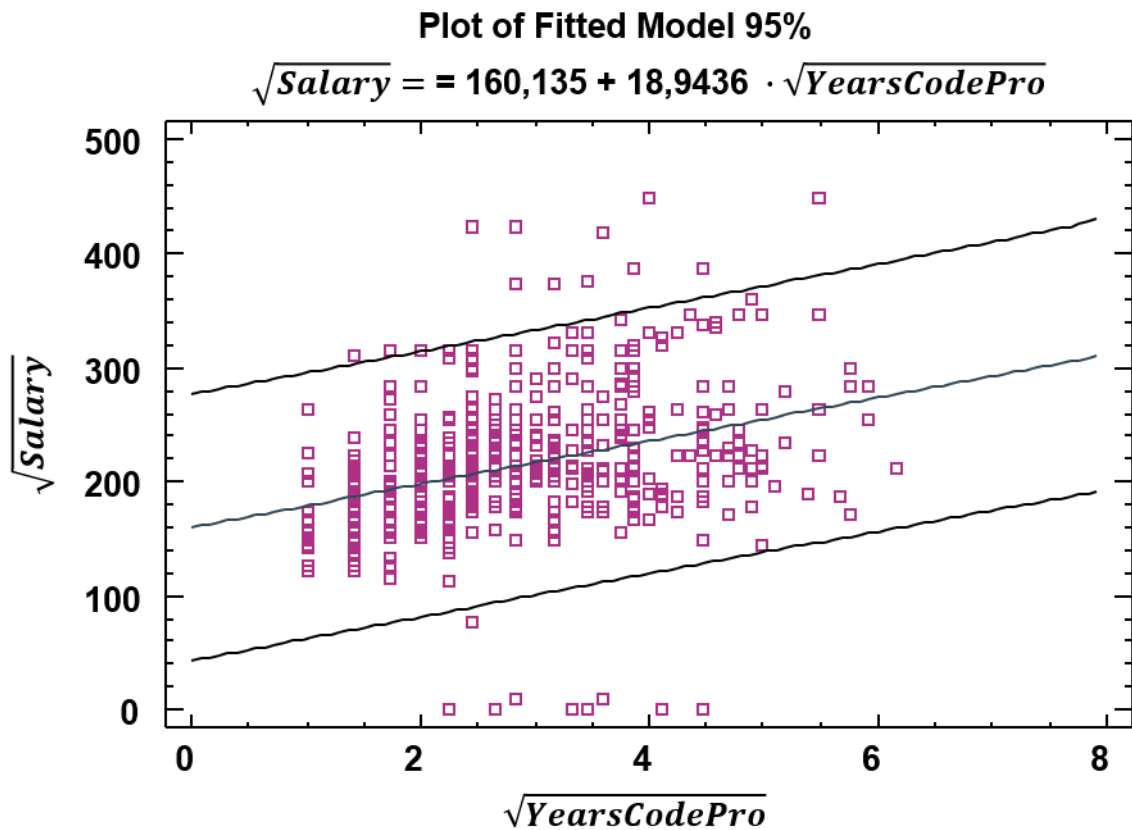
PICTURE 49: Dispersion graph with guidelines.

Since the guidelines tend to converge, that is, they are not parallel, we do not achieve homoscedasticity.

In summary: The relationship is linear, positive and moderate.

**39. Between the quantitative variables, chose the one (Y) that could be considered as response variable. From the matrix of correlation, identify the variable (X) with more correlation with Y. Make a Simple Linear Regression Analysis which allows to predict the values of Y in function of X.**

For this case, even if the variable  $\sqrt{\text{YearsCodePro}}$  has the greatest correlation, the chosen one was  $\sqrt{\text{Salary}}$  because is the most interesting case of study. To predict the value of the salary in function of the years you spent coding as a profession.



PICTURE 50: Plot of dispersion between  $\sqrt{\text{Salary}}$  and  $\sqrt{\text{YearsCodePro}}$  with prediction intervals and regression line.

The prediction interval gives us an idea of how the regression line will be. Basically, we know that the points will be inside the black interval with 95% confidence. They are useful to know how disperse or how close to the line the points will be.

	Estimated LS	Standard Error	T Statistic	P-Value
Intercept	160,135	7,38411	21,6864	0
$\sqrt{\text{YearsCodePro}}$	18,9436	2,35747	8,03559	0

TABLE 31: Regression Model.

Yes, we can say that it is statistically significant since the P-Value of  $\sqrt{\text{YearsCodePro}}$  is 0. It doesn't really matter which  $\alpha$  we chose since the P-Value will be less than  $\alpha$  in all cases. This is coherent with the information we get from the graphic since the majority of the values are inside the limits.

Finally, it is possible to define the equation of the model using the following formula  $Y = A + B \cdot X \rightarrow \sqrt{\text{Salary}} = 160,135 + 18,9436 \cdot \sqrt{\text{YearsCodePro}}$

In this case we have that  $A=160,135$  and  $B=18,9436$ . A is the point of the X axis at which the line intercepts it, while B is the slope of the line. Since the slope is positive, we get that the variables have a positive relation.

Note that the level of significance remains in 0.05.

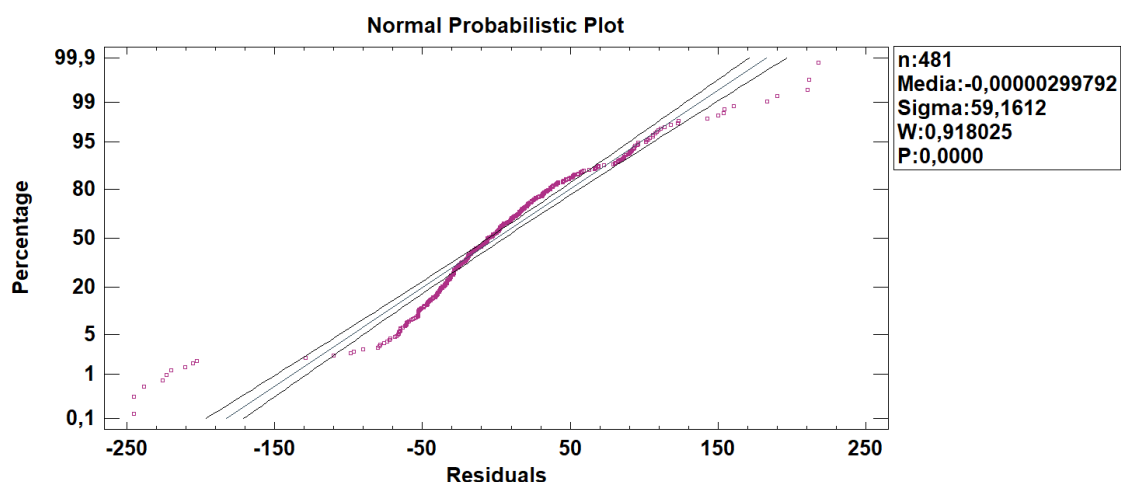
#### 40. With respect to the previous exercise:

We can interpret the coefficients in the following way:

- **Interpretation of A:** When  $\sqrt{YearsCodePro}=0$ , then the square root of the number associated with the Salary is equal to 160,135 on average, or the same, the number associated with the Salary will be  $160,135^2= 25643,21$  on average.
- **Interpretation of B:** When  $\sqrt{YearsCodePro}$  increases in 1, the square root of the number associated with their Salary increases in 18,9436 on average. The number associated with the YearsCode increases because the higher the years the higher the salary, which actually means that it increases. Then, B is the average increase of  $\sqrt{Salary}$  expected if  $\sqrt{YearsCodePro}$  increases one unit.

The correlation observed is due to a cause-effect. The cause is the number of years coding and the effect is the salary, which will be higher the more years coding it has.

#### 41. Save the residuals of the model and represent them on a Normal Probability Plot. What can be deduced?



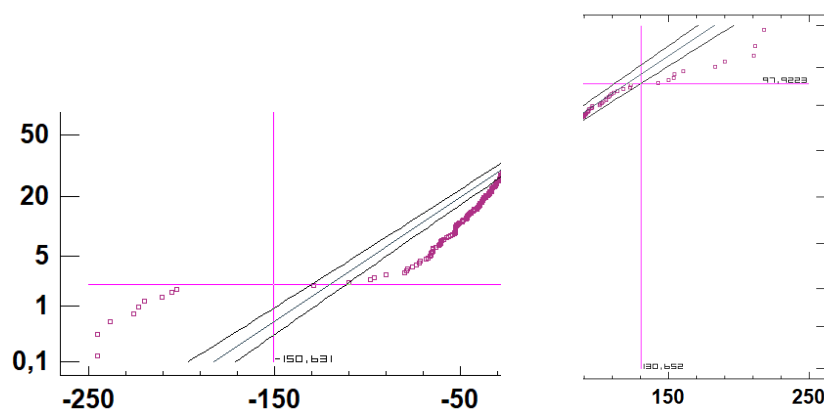
PICTURE 51: Normal Probabilistic Plot for the Model's Residuals.

If the residuals closely follow a straight line, it suggests that the residuals are normally distributed. This indicates that the normality assumption of the model is reasonable, supporting the validity of the statistical inference and assumptions made in the model.

In this case, the data does not follow the line and even surpasses the limits, so we have to reject the normality.

Also it is remarkable the presence of outliers at the extremes. These kind of values should be removed.

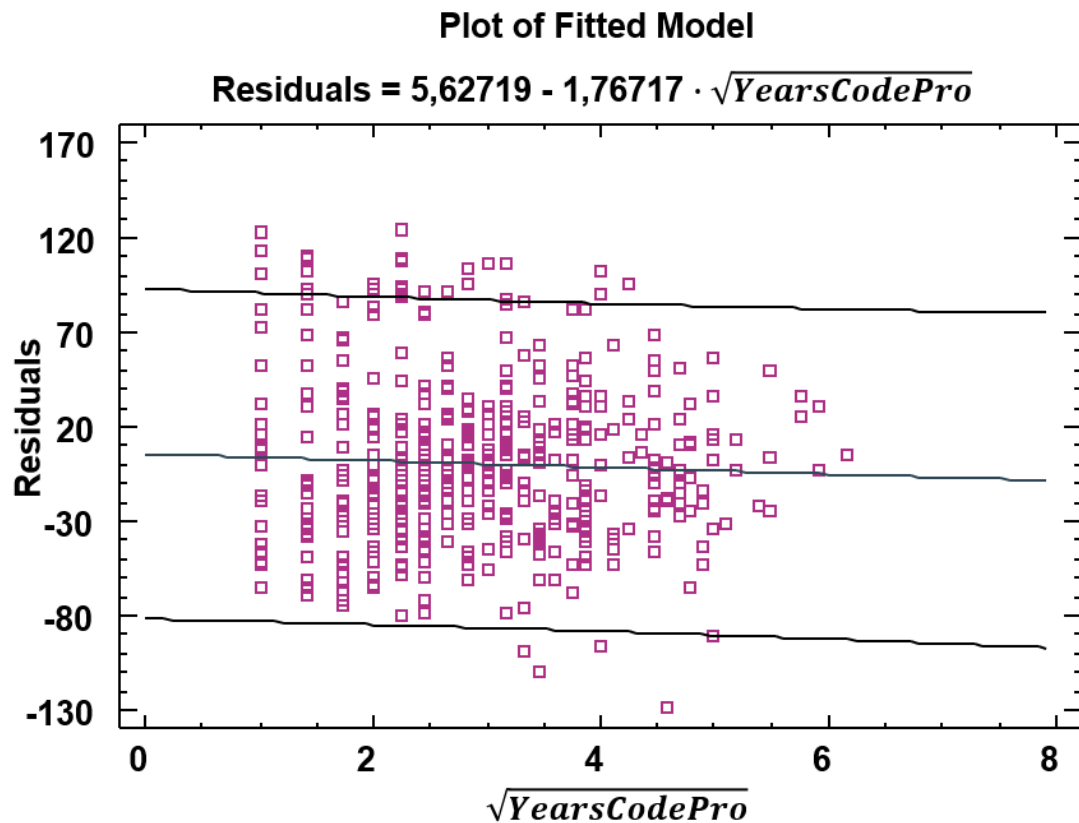
To remove the outliers, some limits were taken using Statgraphics and they were extracted using a Python script (**removeOutliers41.py**).



PICTURE 52,53: Limits taken with Statgraphics

The procedure is the following:

- Move the residuals from Statgraphics to Excel.
- Export the data from Excel to CSV format.
- Use the Python script (or Excel) to take limits and remove the outliers.
- If you used Python, export the data to CSV and load it into Excel again.
- From Excel move the residuals to a new column of Statgraphics.



PICTURE 54: Dispersion Plot of YearsCodePro with Residuals.

Yet again the vast majority of the values are inside the threshold. The dispersion is like a line, which could mean that there is no correlation between  $\sqrt{YearsCodePro}$  and the residuals. Clearly, there is not quadratic effect between both variables. But if we wanted to check it we would need to add a new independent variable which would be YearsCodePro.

In order to check if the effect is statistically significative, let's review the Table of the Model.

	Estimated LS	Standard Error	T Statistic	P-Value
Intercept	5,562719	5,8006	0,970106	0,3325
$\sqrt{YearsCodePro}$	-1,76717	1,84806	-0,956229	0,3395
YearsCodePro	-0,21974	0,290265	-0,757032	0,4494

TABLE 31: Regression Model with Residuals.

As we can see, nor the default value or the squared have a lower P-Value than 0.05. Therefore, we have enough evidence to state that the effect is not statistically significant.



For this particular case, it would be a non sense to try to calculate a prediction interval. As we have just seen, there is not significance. Nevertheless, if we wanted to do it anyways with a quadratic effect, we would have to do the following:

Firstable, we need to compute the mathematical equation of the model by looking at the table. Since this is a quadratic effect, the equation will be of the type

$Y = a + bx + cx^2$ . The values of a, b and c can be seen in the table, in the “Estimate” column. Then, we have that  $a = 5,562719$ ,  $b = -1.76717$  and

$c = -0,21974$  Therefore, the equation of the curve is the following:

$$Residuals = 5,562719 - 1.76717 \cdot \sqrt{YearsCodePro} - 0,21974 \cdot YearsCodePro$$

Then, it is needed to compute the first quartile of X with Statgraphics  $Z_{25}$

Knowing that  $\frac{Residuals}{YearsCodePro} = Z_{75}$  follows a normal distribution, we can compute the average, and we already have the standard deviation.

To calculate the average, we just need to replace YearsCodePro by  $Z_{75}$  on the curve equation.

Finally, we just need to check the values for the given percentiles on a normal distribution in order to get the interval.

This result will tell us that the 95% of the residuals when YearsCodePro is equal to  $Z_{75}$  will be comprised on the previously obtained interval.