Management Degree

Statistics 2

SPSS aplications

Linear regression model with SPSS

2023/2024

NA EXAMPLE OF MULTIPLE LINEAR REGRESSION

We will use the data set **SAMPLE_RLM.sav**, in order to explain the wages of a set of workers in a given activity sector (**Wages**) as a function of the number of years of study (**YearsStudy**), the number of years of specific training (**YearsSpecialization**), the number of years of professional experience (**YearsExperience**).

Some descriptive statistics were obtained:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Wage	70	139	1355	667,17	223,567
Years of Study	70	2	18	9,30	3,200
Years of Specialization	70	0	5	1,70	1,312
Years of professional experience	70	4	24	11,77	4,709
Valid N (listwise)	70				

Degree of linear correlation between quantitative variables under study:

Correlations

		Wage	Years of study	Years of Specialization	Years of professional experience
Wage	Pearson Correlation	1	,946**	,324**	,191
	Sig. (2-tailed)		<,001	,006	,114
	N	70	70	70	70
Years of study	Pearson Correlation	,946**	1	,277*	,183
	Sig. (2-tailed)	<,001		,020	,130
	N	70	70	70	70
Years of Specialization	Pearson Correlation	,324**	,277*	1	,228
	Sig. (2-tailed)	,006	,020		,058
	N	70	70	70	70
Years of professional experience	Pearson Correlation	,191	,183	,228	1
	Sig. (2-tailed)	,114	,130	,058	
	N	70	70	70	70

**. Correlation is significant at the 0.01 level (2-tailed).

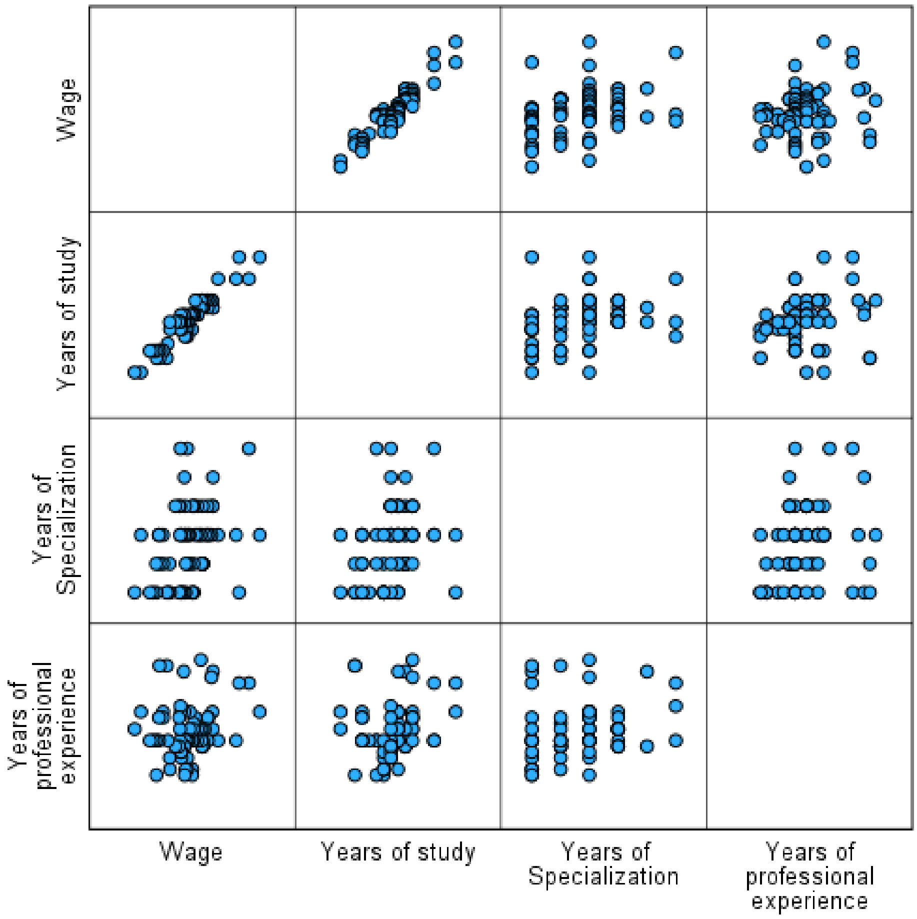
*. Correlation is significant at the 0.05 level (2-tailed).

MULTIPLE LINEAR REGRESSION MODEL

➔ Scatterplot for each pair of variables

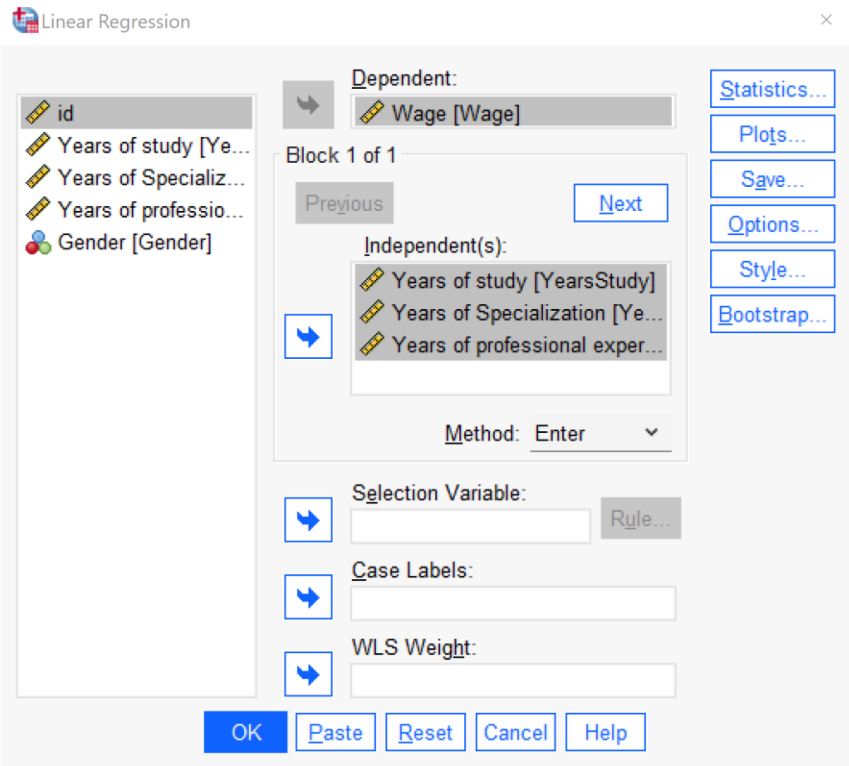
Graphs
Scatter/Dot
Matrix Scatter
Define
Matrix variables: Wage, Years Study, YearsSpecialization, YearsExperience

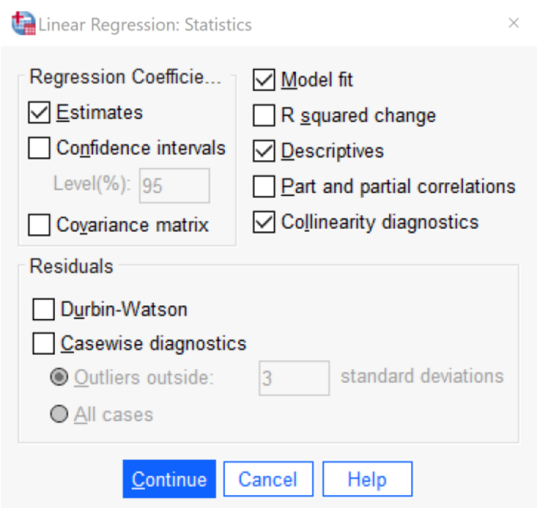
OUTPUTS

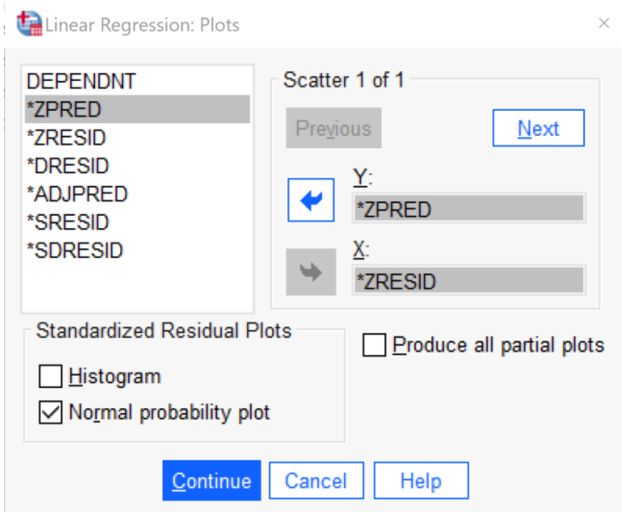


➔ Multiple Linear Regression Model Estimation (forcing all quantitative variables into the model - Enter method)

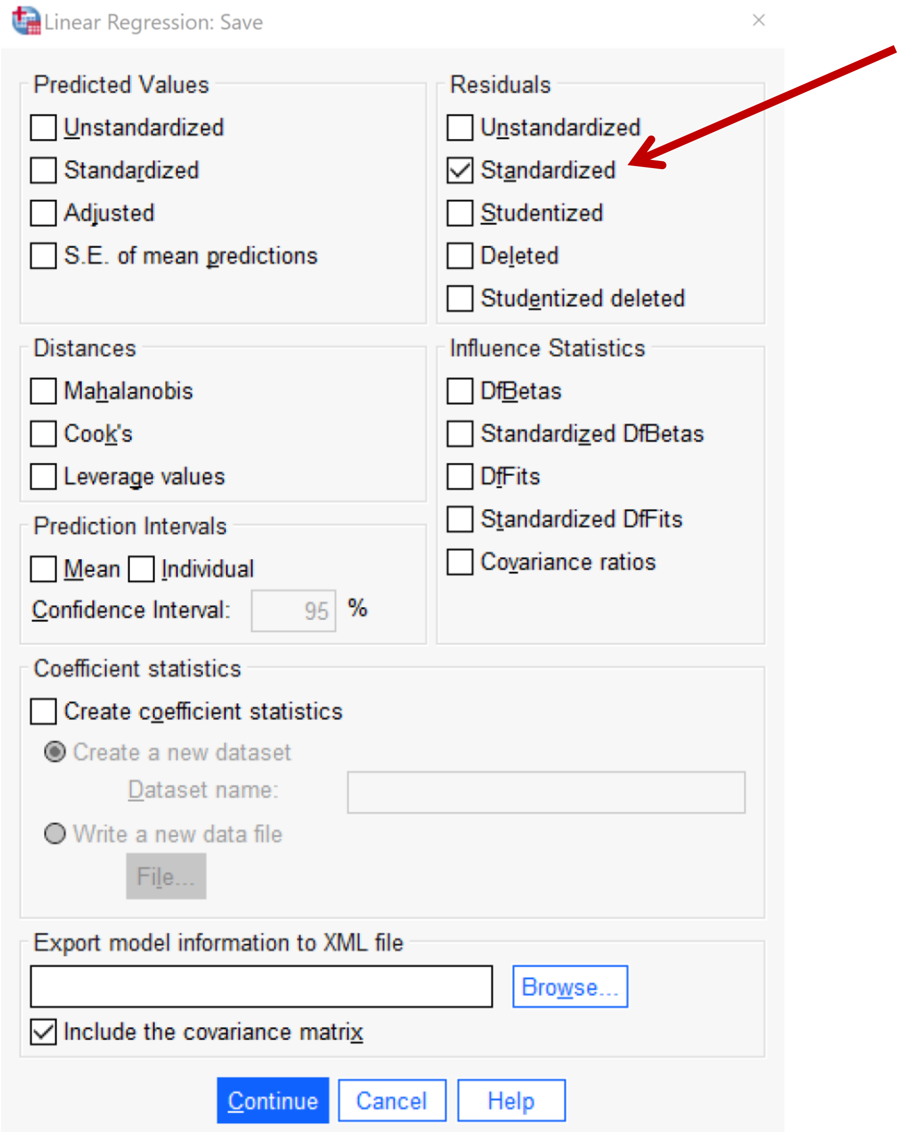
Analize
Regression
Linear







The (standardized) residuals can be saved in the database for later analysis:



OUTPUTS

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Years of professional experience, Years of study, Years of Specialization [b]	.	Enter

a. Dependent Variable: Wage
b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,948[a]	,898	,894	72,909

a. Predictors: (Constant), Years of professional experience, Years of study, Years of Specialization
b. Dependent Variable: Wage

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3097922,903	3	1032640,968	194,261	,000[b]
	Residual	350839,040	66	5315,743		
	Total	3448761,943	69			

a. Dependent Variable: Wage
b. Predictors: (Constant), Years of professional experience, Years of study, Years of Specialization

Coefficients[a]

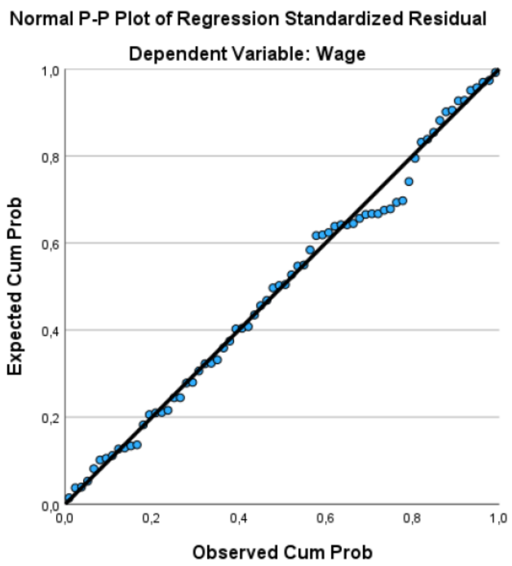
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	42,724	32,202		1,327	,189		
	Years of study	64,708	2,878	,926	22,484	,000	,908	1,101
	Years of Specialization	11,116	7,091	,065	1,568	,122	,891	1,123
	Years of professional experience	,320	1,930	,007	,166	,869	,933	1,072

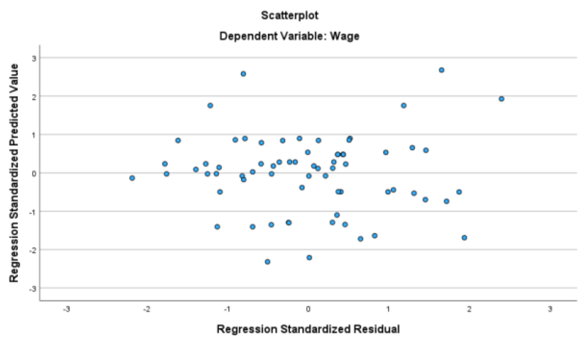
a. Dependent Variable: Wage

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Years of study	Years od specialization	Years of professional experience
1	1	3,587	1,000	,01	,01	,02	,01
	2	,266	3,675	,03	,02	,96	,04
	3	,100	5,976	,02	,36	,00	,74
	4	,047	8,698	,95	,61	,01	,21

a. Dependent Variable: Wage



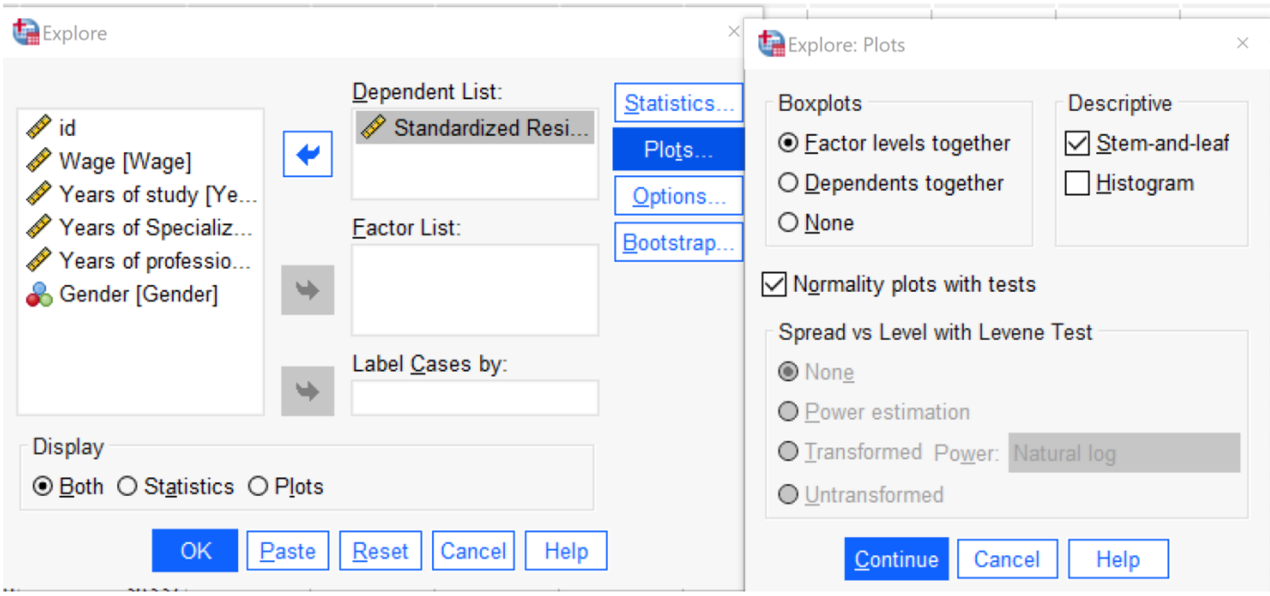


➔ Exploratory analysis of the residuals

Analize

Descriptive Statistics

Explore



OUTPUTS

Descriptives

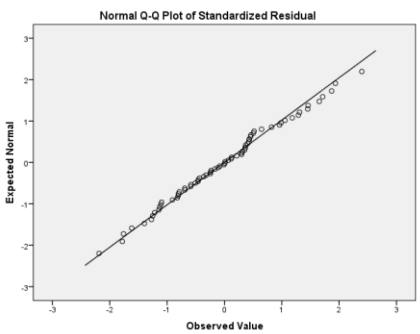
			Statistic	Std. Error
Standardized Residual	Mean		,0000000	,11689566
	95% Confidence Interval for Mean	Lower Bound	-,2332005	
		Upper Bound	,2332005	
	5% Trimmed Mean		-,0082780	
	Median		,0086494	
	Variance		,957	
	Std. Deviation		,97801929	
	Minimum		-2,18698	
	Maximum		2,39583	
	Range		4,58282	
	Interquartile Range		1,18957	
	Skewness		,156	,287
	Kurtosis		-,265	,566

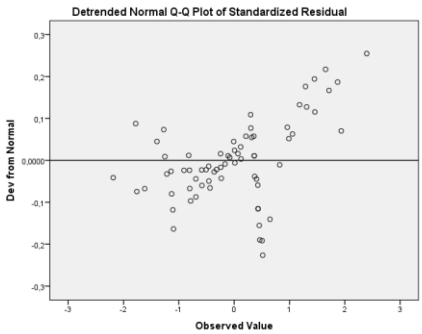
Tests of Normality

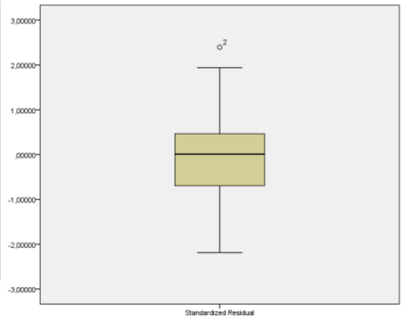
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	,085	70	,200*	,990	70	,857

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction







➔ Interpretation of the results

A. Modelo: $\boldsymbol{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + \varepsilon_i}$ com $\boldsymbol{\varepsilon_i \sim N(0;\sigma)}$

In our case

$$Wage = \beta_0 + \beta_1 YearsStudy + \beta_2 YearsSpecialization + \beta_3 YearsExperience + \varepsilon_i$$

B. Checking the Assumptions:

1. Linear relation between each variable *X* and *Y*

(verify with *scatterplots for* (*Xi, Yi*); correlations analysis)

Correlations

		Wage	Years of Study	Years of Specialization	Years of professional experience
Wage	Pearson Correlation	1	,946**	,324**	,191
	Sig. (2-tailed)		,000	,006	,114
	N	70	70	70	70
Years of Study	Pearson Correlation	,946**	1	,277*	,183
	Sig. (2-tailed)	,000		,020	,130
	N	70	70	70	70
Years of Specialization	Pearson Correlation	,324**	,277*	1	,228
	Sig. (2-tailed)	,006	,020		,058
	N	70	70	70	70
Years of professional experience	Pearson Correlation	,191	,183	,228	1
	Sig. (2-tailed)	,114	,130	,058	
	N	70	70	70	70

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

2. Normality: ε_i follow na approximate normal distribution

(verify with *Normal P-P plot* or Kolmogorov-Smirnov normality test for the residuals)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Standardized Residual	,085	70	,200*	,990	70	,857

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

3. The expected value of the residuals is null: $\mathrm{E}[\varepsilon_i] = 0$

(this assumption **is not verifiable** because the residuals **<u>are estimated in such a way that the sum of the estimates is always zero</u>**, which can be seen in the mean of the Residual variable when calculating the *Residuals Statistics*)

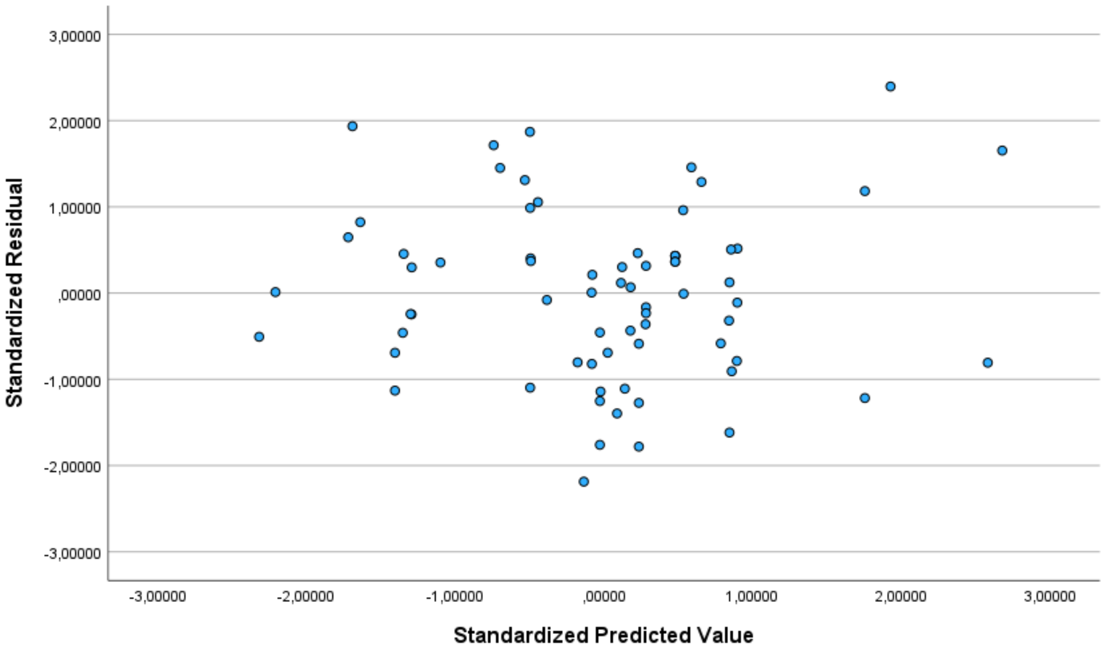
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	175,98	1234,50	667,17	211,890	70
Residual	-159,451	174,678	**,000**	71,307	70
Std. Predicted Value	-2,318	2,677	,000	1,000	70
Std. Residual	-2,187	2,396	,000	,978	70

a. Dependent Variable: Wage

4. Homoscedasticity: The variance of ε_i is constant, i.e., $Var[\varepsilon_i] = \sigma^2$

(check with a *scatterplot* $\left(\widehat{ZY}_i; ze_i\right)$ and see if it increases, or not, the dispersion of the e_i)



5. No autocorrelation: errors are independent, i.e., $Cov(\varepsilon_i; \varepsilon_j) = 0$
(its validation is only relevant if the data are chronological)

6. No multicolinearity: the predictor variables (X_i) are not strongly correlated
(see the correlation matrix; see the *Tolerance* (ok if > 0,1) and *VIF* (ok if < 10), analyze *Condition Indexes* (ok if < 30))

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	42,724	32,202		1,327	,189		
	Years of Study	64,708	2,878	,926	22,484	,000	,908	1,101
	Years of Specialization	11,116	7,091	,065	1,568	,122	,891	1,123
	Years of Experience	,320	1,930	,007	,166	,869	,933	1,072

a. Dependent Variable: Wage

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Years of study	Years of specialization	Years of professional experience
1	1	3,587	1,000	,01	,01	,02	,01
	2	,266	3,675	,03	,02	,96	,04
	3	,100	5,976	,02	,36	,00	,74
	4	,047	8,698	,95	,61	,01	,21

a. Dependent Variable: Wage

C. F-test for global significance of the model (ANOVA table):

H0 : The MLRM **is not** adequate vs H1 : The MLRM **is** adequate

Or

H0: $\beta_1 = \beta_2 = \cdots = \beta_m = 0$ vs H1: $\exists_i : \beta_i \neq 0$, $i=1,2,\ldots,m$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3097922,903	3	1032640,968	194,261	<,001[b]
	Residual	350839,040	66	5315,743		
	Total	3448761,943	69			

a. Dependent Variable: Wage
b. Predictors: (Constant), Years of professional experience, Years of study, Years of specialization

In this case *p-value* < 0.001 < α = 0.05, levando à rejeição da H0. Assim, o modelo é adequado.

D. Percentagem de variância explicada pelo modelo

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,948[a]	,898	,894	72,909

a. Predictors: (Constant), Years of professional experience, Years of study, Years of specialization
b. Dependent Variable: Wage

89,8% of the total variance is explained by the regression

E. Estimação dos parâmetros do modelo:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	42,724	32,202		1,327	,189		
	Years of Study	64,708	2,878	,926	22,484	,000	,908	1,101
	Years of Specialization	11,116	7,091	,065	1,568	,122	,891	1,123
	Years of professional experience	,320	1,930	,007	,166	,869	,933	1,072

a. Dependent Variable: Wage

$$\widehat{Wage} = 42{,}7 + 64{,}7 \times YearsStudy + 11{,}1 \times YearsSpecialization + 0{,}32 \times YearsExperience$$

The estimate of the error dispersion is s=72.9 (identified by Std. Error of Estimate).

Years of Specialization and Years of Professional Experience do not reveal coefficients significantly different from 0 (see *p-values* of t-tests). We would consider redoing the model WITHOUT these variables.

Interpretation of the coefficients

Coefficient	B	
(Constant)	42,724	The estimated average wage, without the influence of the independent variables considered is 42,7 m.u.
Years of Study	64,708	For each additional year of study we expect an increase of 64,7 m.u. in salary, keeping the other variables constant
Years of Specialization	11,116	For each additional year of specific training an increase of 11,1 m.u. in salary is expected, keeping everything else constant
Years of professional experience	,320	For each additional year of work experience a 0,3 m.u. increase in salary is expected, keeping everything else constant

The standardized coefficients (Beta) are used essentially to identify which variables are most influential for the model, in relative terms. The most important explanatory variable is Years of study. These coefficients are interpreted in terms of standard deviations. For example, if $Beta_1 = 0{,}926$, it means that for every one standard deviation increase in years of study, one expects an increase of 0.926 standard deviations in salary.