Regression exercises (from tests)

1.

In a study on consumer buying behavior, one of the objectives was to try to explain the importance of price in the purchase decision. To this end, a multiple regression analysis was performed to try to explain the importance of price (1: not important to 5: very important) as a function of the respondent's age (years), the purchase or not of branded products (1: purchase; 0: non-purchase) and the standardized indicators importance given to brand, importance given to design and importance given to utility. Consider the following results.

Correlations

		Price	*Importance given to the brand*	*Importance given to design*	*Importance given to utility*	*Age*
Pearson Correlation	Price	1,000	-,181	-,003	,751	-,010
	Importance given to the brand	-,181	1,000	-,003	-,007	,025
	Importance given to design	-,003	-,003	1,000	,002	,063
	Importance given to utility	,751	-,007	,002	1,000	-,034
	Age	-,010	,025	,063	-,034	1,000

Model Summary(b)

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
1	,772(a)	,596	,591	,543

a Predictors: (Constant), Buys branded products, Importance given to brand, Importance given to design, Importance given to utility, Age
b Dependent Variable: Price

ANOVA(b)

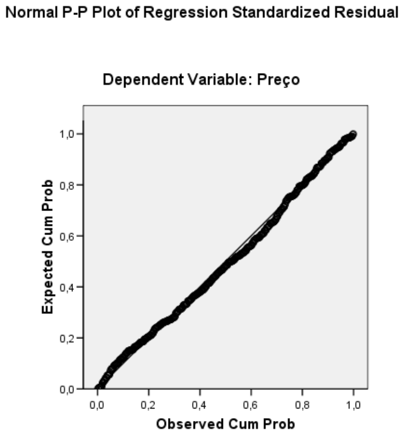
Model		*Sum of Squares*	*Df*	*Mean Square*	*F*	*Sig.(a)*
1	Regression	199,463	5	39,893	135,080	,000
	Residual	135,259	458	,295		
	Total	334,722	463			

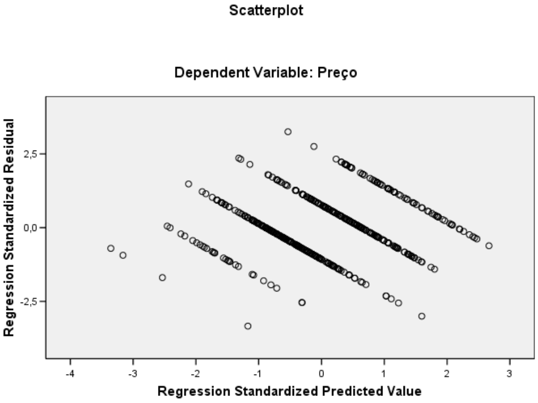
a Predictors: (Constant), Buys branded products, Importance given to brand, Importance given to design, Importance given to utility, Age
b Dependent Variable: Price

Coefficients(a)

Model		*Unstandardized Coefficients*		*Standardized Coefficients*	*t*	*Sig.*	*Collinearity Statistics*	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3,416	,278		12,296	,000		
	Importance given to the brand	-,150	,026	-,178	-5,806	,000	,936	1,068
	Importance given to design	-,006	,025	-,007	-,228	,820	,994	1,006
	Importance given to utility	,637	,025	,752	25,173	,000	,989	1,011
	Age	,009	,013	,021	,717	,474	,989	1,011
	Buys branded products	-,019	,061	-,009	-,307	,759	,924	1,082

a Dependent Variable: Price





a) State all the assumptions underlying the multiple linear regression model. What can you conclude about their verification against the outputs presented?

b) Overall, do you consider the model to be adequate? Justify.

c) Write the estimated model equation.

d) Would you propose the elimination of some independent variables from the model? Justify.

e) Interpret the Unstandardized Coefficient of *the Importance given to the brand*.

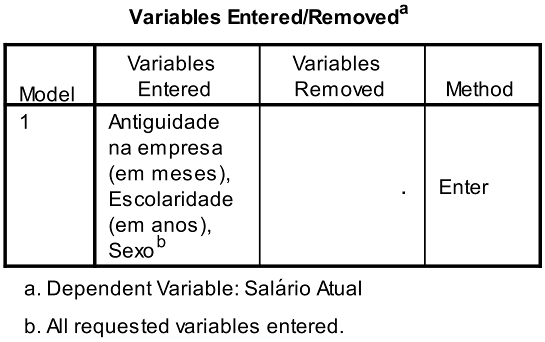
2.

Based on a random sample of 474 employees of an American company, it is desired to find explanatory factors of the current salary (***SA:*** in one thousand m.u.). To this end, the following potentially explanatory factors are considered: the employee's gender (***gender***: 0 – female; 1 – male), education level (***education***, in years), and seniority in the company (***seniority***, in months). Always use a significance level of 0.05.

a) Define the multiple linear regression model for explaining the current salary as a function of the three explanatory variables for the population of all employees of the company.

$$\boldsymbol{SA} = \beta_0 + \beta_1 \times \boldsymbol{gender} + \beta_2 \times \boldsymbol{education} + \beta_3 \times \boldsymbol{seniority} + \epsilon$$

b) The model was estimated using the least squares method, and the following results were obtained, among others:

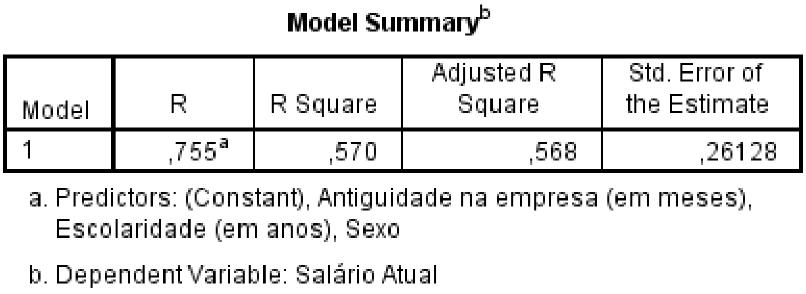


Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Antiguidade na empresa (em meses), Escolaridade (em anos), Sexo[b]	.	Enter

a. Dependent Variable: Salário Atual

b. All requested variables entered.

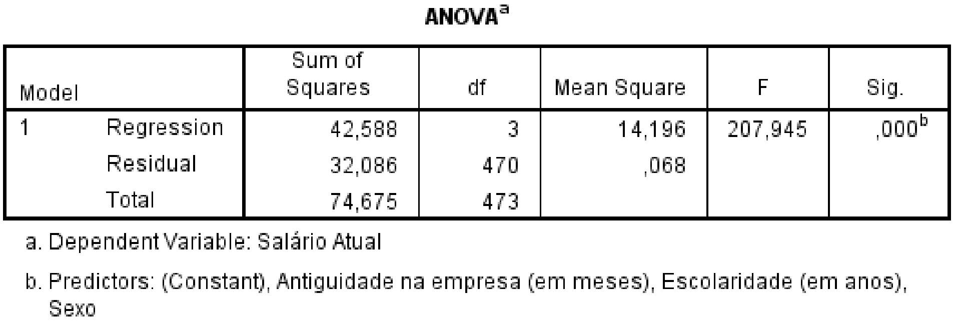


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,755[a]	,570	,568	,26128

a. Predictors: (Constant), Antiguidade na empresa (em meses), Escolaridade (em anos), Sexo

b. Dependent Variable: Salário Atual



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	42,588	3	14,196	207,945	,000[b]
	Residual	32,086	470	,068		
	Total	74,675	473			

a. Dependent Variable: Salário Atual

b. Predictors: (Constant), Antiguidade na empresa (em meses), Escolaridade (em anos), Sexo

b1) Formulate the hypotheses under test and interpret the result obtained.

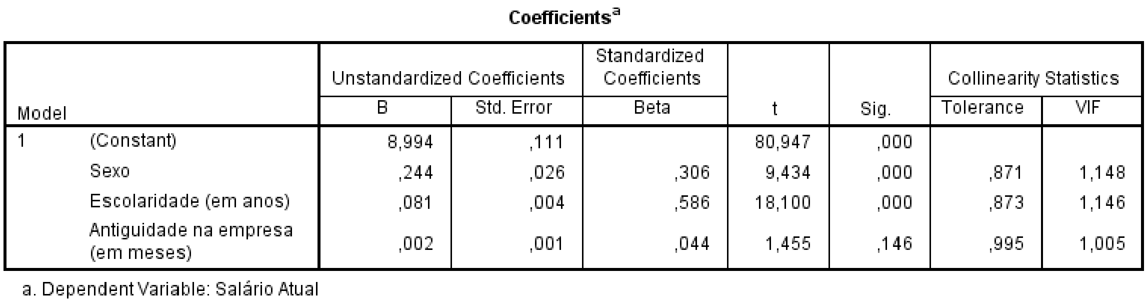
$$\begin{array}{c} H_0\colon \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1\colon \exists\, i : \beta_i \neq 0, i = 1,2,3 \end{array} \quad \text{or} \quad \begin{array}{c} H_0\colon \rho = 0 \\ H_1\colon \rho \neq 0 \end{array}$$

For a significance level of 0.05 as p-value = 0.000, H0 is rejected, so at least one of the independent variables will have a coefficient other than 0, i.e., at least one of the independent variables will explain the current wage (SA).

b2) Interpret the coefficient of determination.

$R^2 = 0.57$ → 57% of the variability of the current salary (AS, dependent variable) is explained linearly by the regression model, i.e., by the variables gender, education and seniority

c) Based on the information below, write the estimated regression line. Interpret the estimates of the coefficients associated with the variable *Gender* and the variable *Sducation*.



Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	8,994	,111		80,947	,000		
	Sexo	,244	,026	,306	9,434	,000	,871	1,148
	Escolaridade (em anos)	,081	,004	,586	18,100	,000	,873	1,146
	Antiguidade na empresa (em meses)	,002	,001	,044	1,455	,146	,995	1,005

a. Dependent Variable: Salário Atual

$$\widehat{SA} = 8.994 + 0.244 \times \boldsymbol{gender} + 0.081 \times \boldsymbol{education} + \ 0.002 \times \boldsymbol{seniority}$$

Gender: Being male, compared to being female, have an increase of 0.244 m.u. in the SA (*ceteris paribus*).

Education: For each additional year of education, it's expected an increase of 0.081 m.u. in the SA (*ceteris paribus*).

d) Estimate the average salary of a male worker with 12 years of education, at the time he is hired by the company?

$$\widehat{SA} = 8.994 + 0.244 \times \mathbf{1} + 0.081 \times \mathbf{12} + \ 0.002 \times 0 = 10{,}21 \text{ m.u.}$$

e) Indicate, justifying, which are the independent variables with a significant impact on the current salary. Formulate the hypotheses under test generically.

We can analyse the t-test results, assuming a significance level of 5%, for the Hypotheses (generically written)

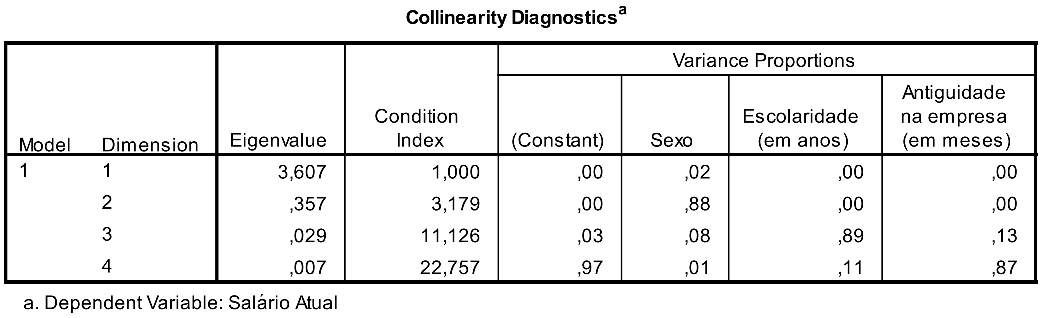
$$H_0\colon \beta_i = 0$$
$$H_1\colon \beta_i \neq 0, i = 1, 2, 3$$

Gender: p-value=0.000 ≤ α=0.05, we Reject H0, so the variable Gender significantly linearly explain SA;

Education: p-value=0.000 ≤ α=0.05, we Reject H0, so the variable Education significantly linearly explain SA;

Seniority: p-value=0146 > α=0.05, we Do not Reject H0, so the variable Seniority does not have a significant impact in linearly explaining SA;

f) Study the existing multicollinearity in the model taking into account the available information (previous and following tables)



Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Sexo	Escolaridade (em anos)	Antiguidade na empresa (em meses)
1	1	3,607	1,000	,00	,02	,00	,00
	2	,357	3,179	,00	,88	,00	,00
	3	,029	11,126	,03	,08	,89	,13
	4	,007	22,757	,97	,01	,11	,87

a. Dependent Variable: Salário Atual

1. **Tolerance/VIF:** for all independent variables, there are no values lower than 0.1 (>10 in the case of VIF), so there are no multicollinearity problems.

2. **Condition Index:** as the highest Condition index is less than 30, it is confirmed that there is no multicollinearity.

3.

Based on a European survey on working conditions, some indicators were constructed, which are intended to be used to explain the individual's net monthly income (REND: What is the net monthly income from your work?):

Indicator1	Incidence of sedentary work (%)
indicator2	Incidence of physically demanding work (%)
indicator3	Incidence of work in contact with the public (%)
indicator4	Intensity of feeling of autonomy (%)
indicator5	Intensity of feeling of usefulness (%)
indicator6	Intensity of feeling stressed (%)
indicator7	Intensity of feeling of support (%)
indicator8	Intensity of feeling enough time to complete tasks (%)
indicator9	Percentage of time your work depends on others

A first analysis led to the following outputs:

ANALYSIS 1.

Table 1. Descriptive Statistics

	Mean	Std. Deviation	N
REND: What is the net monthly income from your work?	736.71	412.68	162
Q18. Working hours per week	38.88	11.13	162
Incidence of sedentary work (%)	33.24	26.39	162
Incidence of physically demanding work (%)	39.68	19.95	162
Incidence of work in contact with the public (%)	32.90	24.88	162
Intensity of feeling of autonomy (%)	50.77	22.89	162
Intensity of feeling of usefulness (%)	77.29	15.50	162
Intensity of feeling stressed (%)	43.31	18.35	162
Intensity of feeling of support (%)	64.59	14.08	162
Percentage of time your work depends on others	2.97	1.90	162

Table 2. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.623	0.388	0.352	332.308

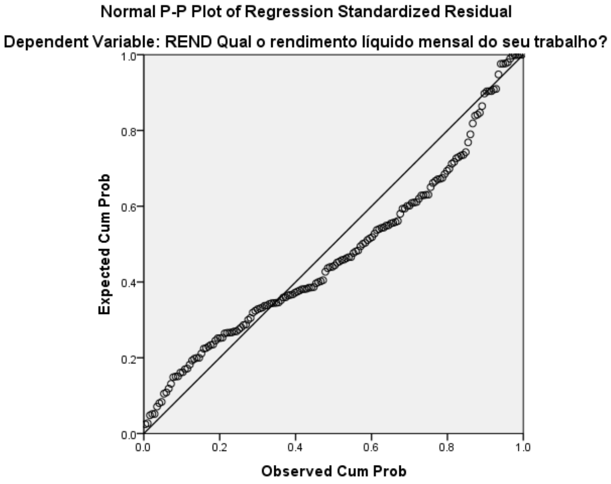
Table 3. ANOVA

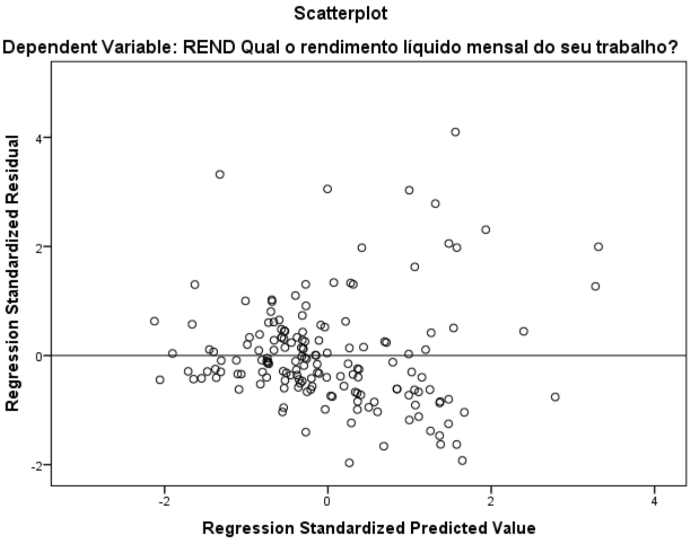
Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	10633701.741	9	1181522.416	10.699	.000
	Residual	16785161.624	152	110428.695		
	Total	27418863.364	161			

Table 4. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-164.247	218.564		-0.751	0.454		
	Q18. Working hours per week	12.190	2.514	0.329	4.848	0.000	0.876	1.142
	Incidence of sedentary work (%)	5.548	1.097	0.355	5.060	0.000	0.819	1.221
	Incidence of physically demanding work (%)	-3.653	1.462	-0.177	-2.499	0.014	0.806	1.240
	Incidence of work in contact with the public (%)	4.696	6.079	0.283	0.773	0.441	0.030	33.357
	Intensity of feeling of autonomy (%)	3.213	1.190	0.178	2.700	0.008	0.924	1.082
	Intensity of feeling of usefulness (%)	1.431	1.699	0.054	0.842	0.401	0.990	1.011
	Intensity of feeling stressed (%)	1.556	1.611	0.069	0.966	0.336	0.785	1.274
	Intensity of feeling of support (%)	0.939	1.970	0.032	0.477	0.634	0.892	1.122
	Percentage of time your work depends on others	-56.914	79.410	-0.262	-0.717	0.475	0.030	33.181

a. Dependent Variable: REND What is the net monthly income from your work?





a) Write down all the MRLM assumptions and, according to the information given, **explain** which are or are not validated and why.

Assumptions of the Multiple Linear Regression Model:

i) Linearity of the phenomenon under study: the relationship between each of the independent variables, Xi, and the dependent variable, Y, must be considered linear;

ii) The residual variables εi have a Normal distribution;

iii) The residual variables εi have a null mean distribution;

iv) The residual variables εi have a distribution with equal variance (homoscedasticity);

v) The residual variables εi are independent, i.e., there is no autocorrelation between residuals; and

vi) No multicollinearity, i.e., the independent variables should not be correlated

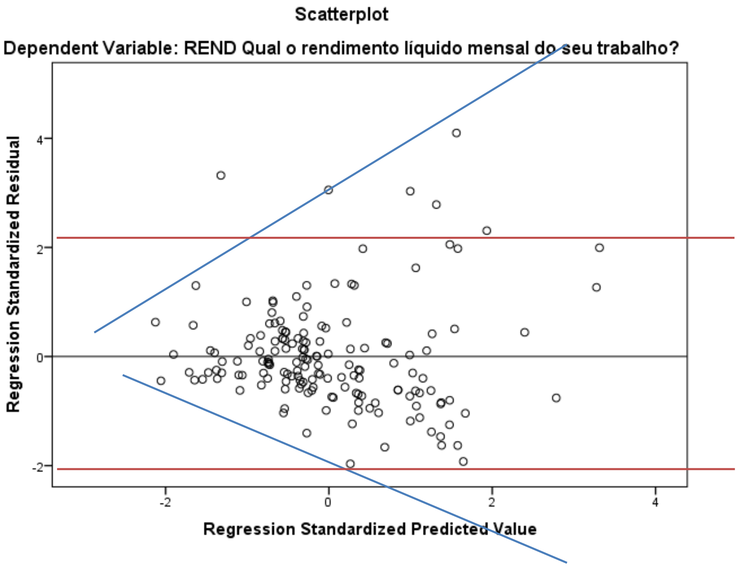
Regarding the validation of these assumptions:

i) To validate this assumption it would be necessary to have the graphs (Xi, Y), i=1 ... 9, or the values of the correlation matrix, which does not happen → we do not have information to validate this assumption

ii) From the "Normal P-P plot" presented we can get an idea of the normality of the errors. In this case, there are some deviations from normality (if there are no other problems, these deviations, although systematic, do not appear to be very serious)

iii) It is not necessary to validate the nullity of the residual average, since the estimated residuals have a null average per construction

iv) This assumption is graphically validated by the analysis of the graph (Zpred; Zresid); In particular, "funnel" patterns are sought, which reveal a systematic increase/decrease in the variance of the residuals. By the analysis of the said graph, there may be some lack of variance homogeneity (although the "funnel" marked below leaves a dot out)



v) This assumption is also graphically validated by the analysis of the graph (Zpred; Zresid); It will be considered validated if the residues are regularly distributed above and below the axis of the XX's, in a "band". In this case (see the red lines above), almost all points are in the approximate band (-2; 2) so there should be no problems regarding the independence of residuals.

vi) The existence or not of strong multicollinearity is made through the analysis of the VIF (or the tolerance coefficients). Two variables present VIF values well above the maximum usually mentioned (10): "Incidence of work in contact with the public" (VIF = 33.4) and "Percentage of time in which their work depends on third parties" (VIF = 33.2)

In conclusion, the model that was intended to be adjusted reveals flaws in the application assumptions.

b) Which explanatory variables are candidates to leave the model and why?

First of all, at least one of the variables with high VIF, and already mentioned in the previous paragraph (as only two have high VIF, it is to be suspected that these are the strongly related ones). In any case, both the "Incidence of working in contact with the public" (VIF = 33.4) and the "Percentage of time in which your work depends on third parties" (VIF = 33.2) also have significance in the corresponding t-test above the reference value α= 0.05 (sig = 0.441 and 0.475, respectively), so that the respective coefficients are not significantly different from 0, therefore, also for this reason, they are both candidates to leave the model.

Other variables also reveal coefficients not significantly different from zero, and are therefore candidates to leave the model: "Intensity of feeling of usefulness" (sig = 0.401); "Intensity of feeling stressed" (sig=0.336) and "Intensity of feeling of existence of support" (sig=0.634)

A second analysis was carried out and led to the following results:

ANALYSIS 2.

Table 5. ANOVA

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	10319507.152	4	2579876.788	23.687	.000b
	Residual	17099356.212	157	108913.097		
	Total	27418863.364	161			

a. Dependent Variable: REND What is the net monthly income from your work?

Table 6. Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1 (Constant)	-13.576	118.386		-0.115	0.909		
Q18. Working hours per week	13.056	2.392	0.352	5.457	0.000	0.954	1.048
Incidence of sedentary work (%)	5.842	1.033	0.374	5.656	0.000	0.911	1.098
Incidence of physically demanding work (%)	-2.835	1.324	-0.137	-2.142	0.034	0.970	1.031
Intensity of feeling of autonomy (%)	3.170	1.181	0.176	2.684	0.008	0.925	1.081

a. Dependent Variable: REND What is the net monthly income from your work?

c) Test the suitability of the model

Hypotheses under test: H_0: $R^2 = 0$ (unsuitable model) vs H_1: $R^2 \neq 0$ (suitable model)

These assumptions are the ones underlying the ANOVA Table.

Since sig = 0.000 < α = 0.05 then we reject H_0, i.e. we can conclude that the model is adequate

d) Write the equation of the estimated model.

Average Net Monthly Income = - 13.576

+ 13,056 Working Hours per week

+ 5,842 Incidence of sedentary work

- 2,835 Incidence of physically demanding work

+ 3,170 Intensity of feeling of autonomy

e) Interpret the **estimated non-standardized coefficient** for the variable "q18. Working hours per week".

B_1 = 13,056

For each additional hour of work per week, the average salary rises by 13,056 euros (or m.u., in the statement it is not explicit, both were accepted), keeping everything else constant.

f) Interpret the standardized coefficient **estimated** for the variable "Incidence of physically demanding work (%)"

$BETA_3$ = -0.137

It is the LEAST important variable for the model, because |BETA3| = 0.137 < |BETAj|, j ≠3

In addition, it can also be interpreted incidentally as follows: for each additional standard deviation in the incidence of physically demanding work, the average monthly net income DECREASES 0.137 standard deviations (keeping everything else constant).

(note that the increase of 1 standard deviation in this variable corresponds to 19.95 pp. By the usual reading of the respective non-standardized coefficient, this implies a variation of 19.95 * (-2.835) € = -56.558 €; as the standard deviation of income is 412.68€, the ratio -56.558/412.68 = -0.137 gives us the variation of income in terms of standard deviations)

g) How much do you predict that the average salary of a worker who works 40 hours a week will be, has an incidence of sedentary work of 80%, an incidence of physically demanding work of 10% and an intensity of feeling of autonomy of 80%?

Estimated average monthly income = - 13,576

+ 13.056 x 40

+ 5.842 x 80

- 2.835 x 10

+ 3.170 x 80 = 1201.27 €