**Correlation and regression analysis**

**Introduction**

|  |
| --- |
| Carrying out a study in the medical field, involves, as seen in previous laboratories, the development of a research protocol. Among other things, it involves identifying variables of interest and establishing the statistical methods needed to test the objectives of the study.  Correlation analysis and regression analysis are some of the most common statistical methods used in studies (scientific articles) in the medical field. |

## Purpose and utility of the laboratory

* Explain when it is appropriate to use:
  + linear correlation analysis
  + linear regression
  + logistic regression
* Acquiring the necessary skills for the critical assessment of studies (scientific articles) in relation to the presented methods

## Linear correlation analysis: purpose, applicability conditions, interpretation of results

**Aim:**

* Linear correlation analysis is used when the problem of interest is to evaluate the intensity of the linear link / relationship between two quantitative variables.

**Parameters:**

* Pearson coefficient of correlation r
* Coefficient of determination d=r2

In this linear relationship analysis

* + we have neither dependent variables (result variables) nor independent variables.
  + there is no intention to predict the values of one of the variables according to the other.
  + the two variables can be interchanged without affecting the value of the correlation coefficient.

**Data types that can be used**

* quantitative

**Applicability conditions**

* the observations are independent of each other (patients are not related, e.g. twins)
* quantitative variables follow a normal probability distribution
* no extreme values

**Interpretation of the linear correlation coefficient (Pearson correlation coefficient)**

* As the correlation coefficient approaches 0, the intensity of the linear link between the two variables decreases;
* As the correlation coefficient approaches 1 or -1, the intensity of the linear link between the two variables increases;
* The correlation coefficient can take values between -1 and +1, and its sign means the type of linear link: the sign - indicates indirect linear relationship (negative correlation, when one variable increase the other decrease), while the + sign indicates a direct link (positive correlation, when one variable increase the other increase too).
* Colton Rules [Colton T. Statistics in Medicine. Little Brown and Company, New York, NY 1974]:
  + r in [-0.25, +0.25] → No relation
  + r in (0.25, +0.50] or in (-0.50, -0.25] → Weak relation
  + r in (0.50, +0.75] or in (-0.75, -0.50] → Moderate relation
  + r in (0.75, +1] or in [-1, -0,75) → Strong relation

!!!!! If neither of the two variables is normally distributed (or the data cannot be transformed to obtain normal distributions) then the Spearman correlation coefficient is used (which will evaluate the intensity of the monotone relationship between the two variables). Spearman correlation coefficient can be interpreteted with the same Colton rules.

*Monotonic relationship* means that when one variable increases the other increases too or when one variable increases the other decreases, but not neccesary in a linear relationship.

## Linear correlation analysis: Scenario 1

An analytical, observational, and cross-sectional study was performed in male patients aged 18 to 30 years, diagnosed with idiopathic hypogonadotropic hypogonadism (HHI) in order to investigate the linear relationship between testosterone levels (nmol / L) and osteocalcin level (protein hormone with an important role in bone mineralization, ng / ml).

The study used a consecutive sample of 88 male subjects aged 18-30 years, patients recruited from the outpatient department of the Department of Endocrinology, Diabetes, Nutrition and Metabolic Diseases of a clinic.

Testosterone and osteocalcin levels were determined from each patient's blood sample.

The histogram of each variable was performed and it was observed that both osteocalcin and testosterone were normally distributed.

**1. Specify the type of data collection (control case / exposed / unexposed / representative sample)**

|  |
| --- |
|  |

**2. Specify the name and type of each study variable (dichotomous qualitative, nominal qualitative, ordinal qualitative, continuous quantitative)**

|  |
| --- |
| testosterone levels (nmol / L) ...  osteocalcin level (ng / ml) ... |

**3. Specify the name of the statistical method appropriate to the data in this study**

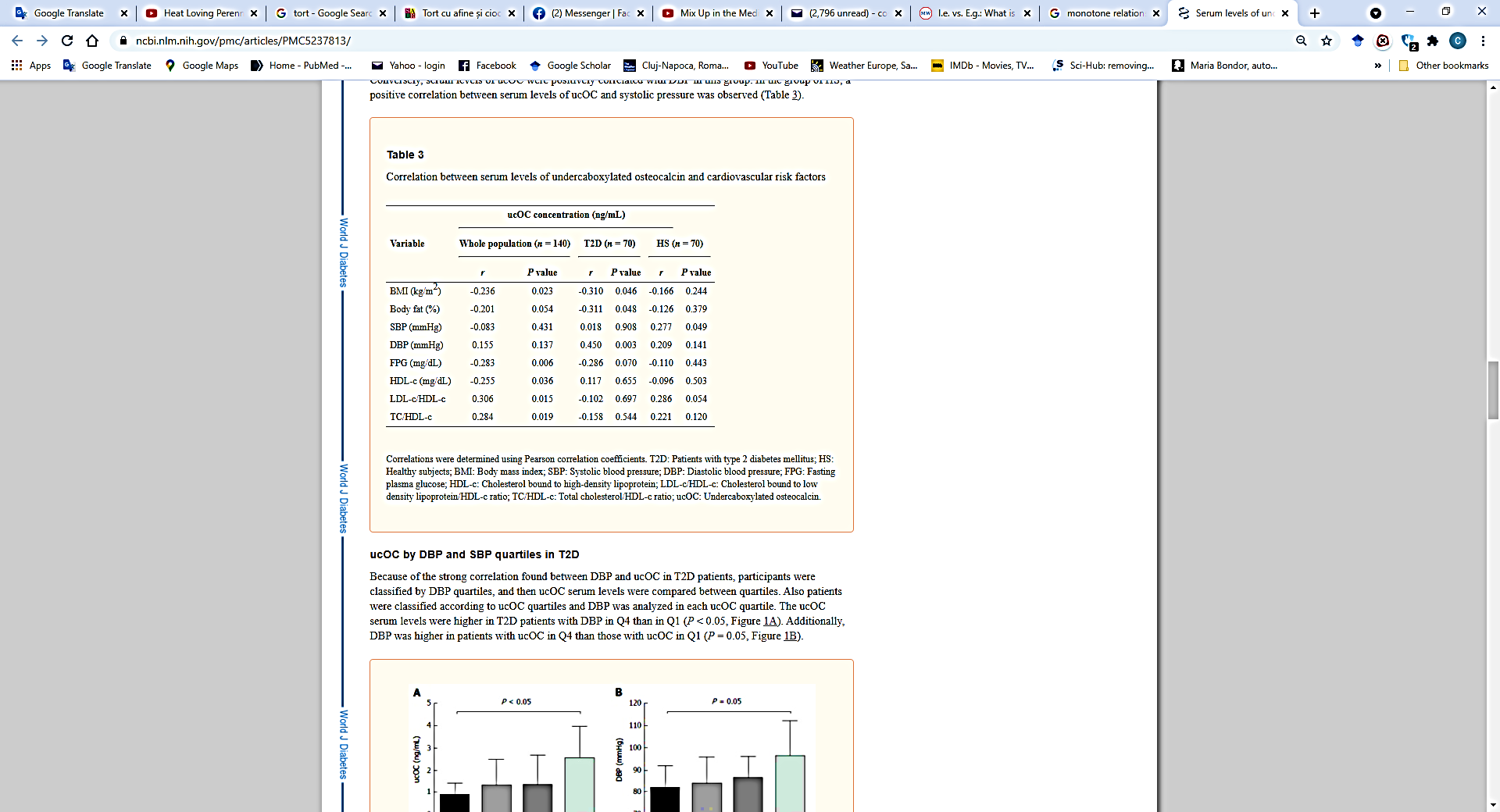
|  |
| --- |
|  |

## Linear correlation analysis: Scenario 2

You want to conduct a cross-sectional observational study to assess the relationship between undercaboxylated osteocalcin (ucOC) (ng/ml) levels and cardiovascular risk factors in adult patients with type 2 diabetes and in healthy subjects.

After a search in PubMed you find an article published on this topic: Sanchez-Enriquez et al [[[1]](#footnote-1)].

A sample of 140 subjects was chosen: 70 patients with type 2 diabetes (T2D) and 70 healthy subjects (HS) who participated in a program for the detection and treatment of congenital and acquired metabolic diseases of a university center.



**1. Specify the type of data collection (control case / exposed - unexposed / representative sample)**

|  |
| --- |
|  |

**2. Specify the name and type of each study variable (dichotomous qualitative, nominal qualitative, ordinal qualitative, continuous quantitative) - see table 3 in the Results section [1]**

|  |
| --- |
| BMI(kg/m2) ...  Body fat (%)  SBP (mmHg)  DBP (mmHg)  FPG (mg/dL)  HDL-c (mg/dL) |

**3. Specify the name of the statistical method appropriate to the data in this study**

|  |
| --- |
|  |

**4. Choose the correct interpretation of coeficient of correlation between DBP and ucOC r=0.450, p=0.003**

|  |
| --- |
| a. Weak positive relation  b. Moderate positive relation  c. Strong positive relation  d. No relation |

**4. Choose the correct interpretations of p assosiated with the coeficient of correlation between DBP and ucOC r=0.450, p=0.003**

a. there is a statistically significant corelation between DBP and ucOC in population

b. there is not a statistically significant corelation between DBP and ucOC in population

c. there is a statistically significant corelation between DBP and ucOC in the sample

d. there is not a statistically significant corelation between DBP and ucOC in the sample

e. if we repeat the study we have more than 95% chances to found an r=0.45 or near.

## Linear correlation analysis: Scenario 3

You want to do a study to see if there is a relationship between the level of leptin (a substance synthesized by adipocytes) and body mass index (BMI).

After a search in PubMed you find an article published on this topic: Bahathiq et al [[[2]](#footnote-2)].

The study was conducted on a sample of 240 women aged between 18 and 65 years. These were divided into three groups according to the value of the body mass index (BMI): a control group (n1 = 80, 18≤IMC≤29.99), a group with obesity (n2 = 80, BMI≥30) and a group with diabetes and obesity (n3 = 80). Leptin levels were measured from each patient's blood sample.

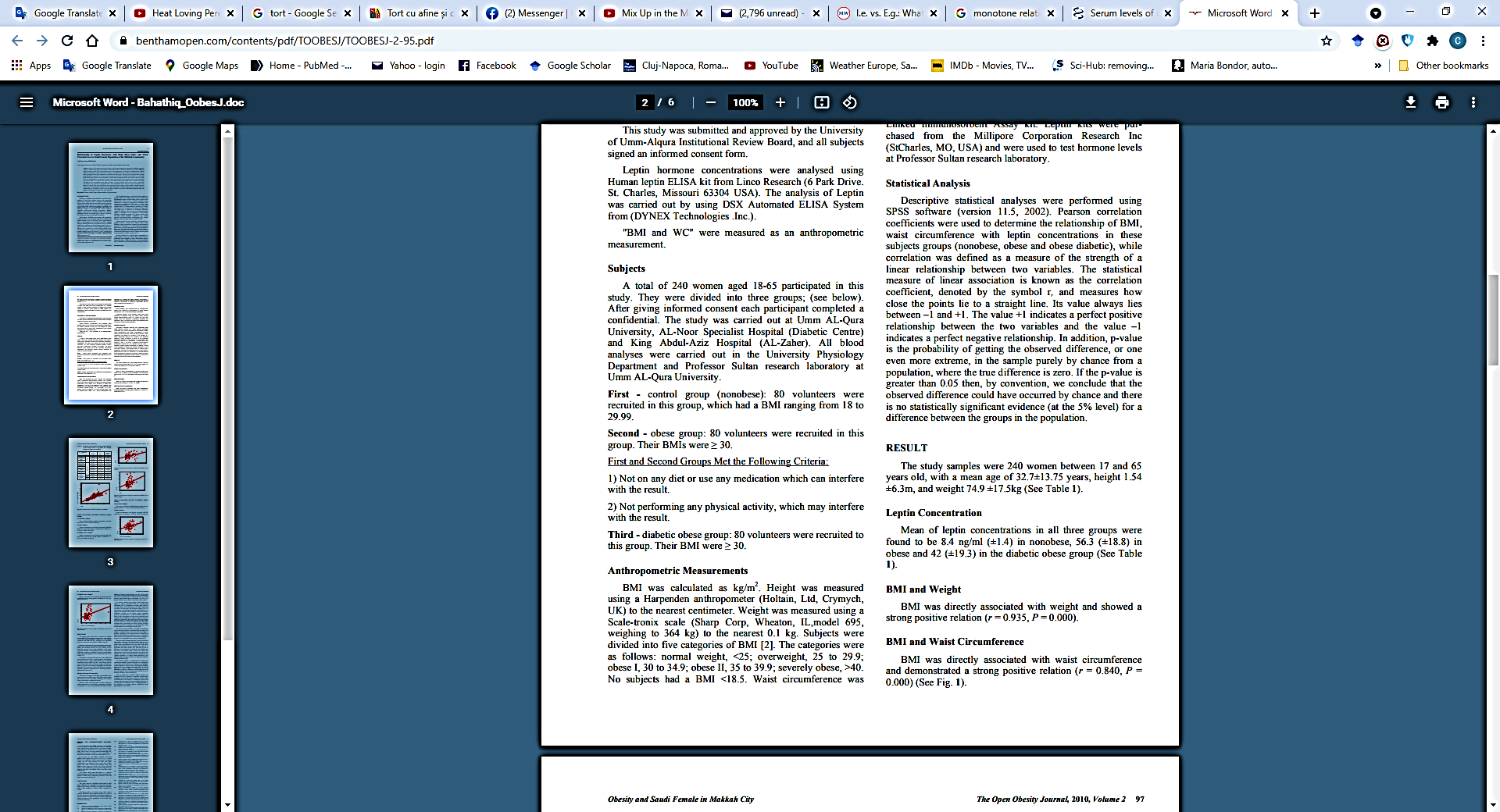
The variables of interest were: BMI (kg / m2), leptin (ng / ml).

1. **Specify the name and type of each variable of interest (dichotomous qualitative, nominal qualitative, ordinal qualitative, continuous quantitative)**

|  |
| --- |
| BMI (kg / m2)...  leptin (ng / ml) |

1. **In the Statistical Analysis section of the article**

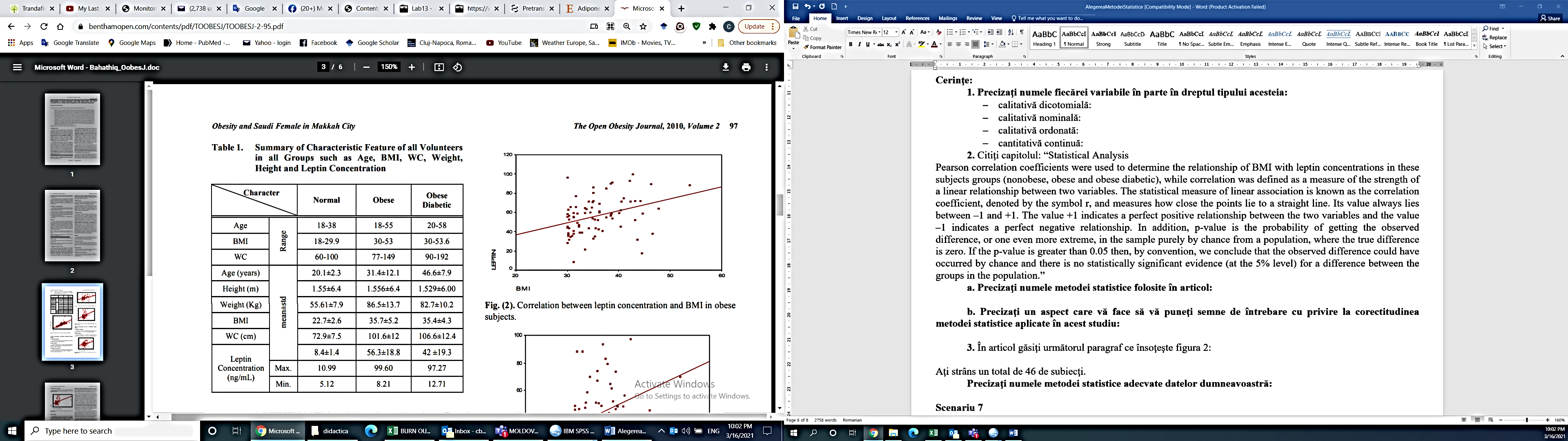
<https://benthamopen.com/contents/pdf/TOOBESJ/TOOBESJ-2-95.pdf>**:**



**Specify an aspect that makes you question the correctness of the statistical method applied in this study:**

|  |
| --- |
|  |

1. **In the article you will find the following paragraph accompanying Figure 2: “In Obese Subjects Leptin concentrations were directly associated with Body Mass Index and demonstrated a strong positive relationship (r = 0.350, p = 0.001) (See Fig. 2).” [2]**

[2]

**Specify the following aspects:**

**a. According to you which is the range with the most BMI values? 30-35, 35-40, 40-45 or> 45?**

|  |
| --- |
|  |

**b. What is the approximate average BMI as can be seen from the graph?**

|  |
| --- |
|  |

**c. Do you assess whether BMI has a normal distribution?**

|  |
| --- |
|  |

**d. Do you consider BMI to have extreme values?**

|  |
| --- |
|  |

**6. Conclusion: (concerning the correctness of the method applied. Justify the answer)**

|  |
| --- |
|  |

## Correlation analysis: Scenario 4

Once you get suspicions about the correctness of the analysis made in the previous article, you decide to find an article that comes from a better rated journal. You find: Elhafeez et al [[[3]](#footnote-3)]. Recognized Springer Publishing House recommends this article, however the journal is not ISI indexed.

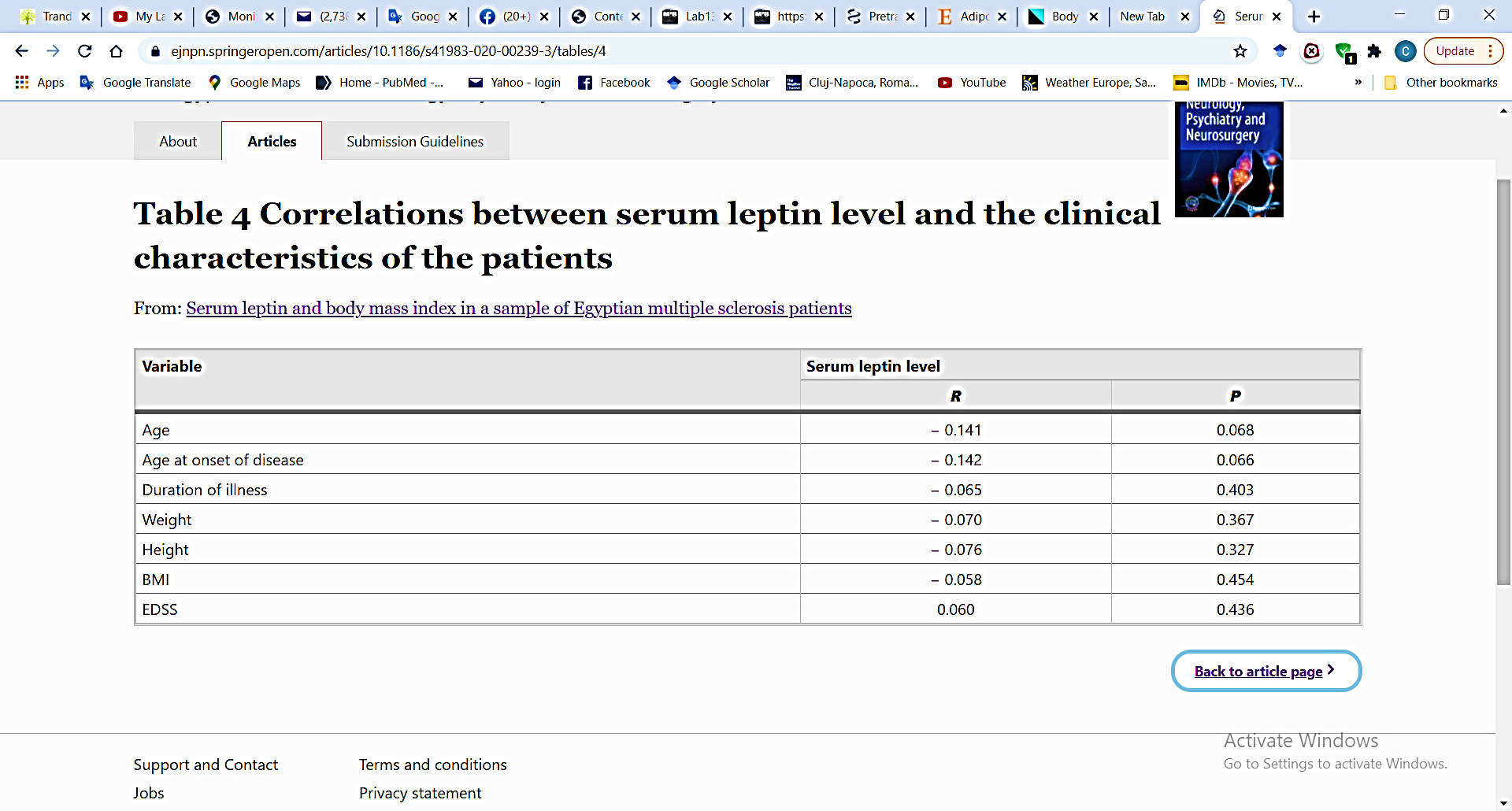
In this study, 169 consecutive patients with multiple sclerosis and a control group of 50 subjects were selected. Patients were evaluated and a blood sample was taken from which leptin was determined.

In the "Statistical methods" chapter of the article, a statistical method used for correlation analysis draws your attention: the Sperman correlation coefficient. „The correlation analysis (using Spearman’s method) was used to assess the strength of association between two quantitative variables. The correlation coefficient denoted symbolically “rho” defines the strength and direction of the linear relationship between two variables. ” [3]

**1. Point out the inaccuracy between this statement and what you know from theory:**

|  |
| --- |
|  |

**2. Although you have found an inaccuracy, you decide to continue reading, the statistical method applied is still good. Find Table 4 and the associated text: "There were no significant correlations between serum leptin level and age of the patient, age at onset of disease, duration of illness, BMI, or EDSS score (Table 4)." [3]**

[3]

**Interpret the Sperman correlation coefficient ρ = -0.058, p = 0.454 between leptin and BMI.**

|  |
| --- |
|  |

## Linear regression analysis: purpose, applicability conditions, interpretation of results

**Purpose** :

* identifies the shape and direction of the linear relationship between one or more explanatory variables (independent variables / predictors) and a dependent variable (result variable / response variable) of quantitative type
* allows us to estimate the effects of several independent variables on a dependent variable:
  + to test hypotheses about predictive factors
  + to produce / create / develop a predictive model.

**Data types that can be used**:

*Dependent variable* Y

* + quantitative

*Independent variable* X1, X2,….,Xk

* + quantitative
  + qualitative
    - dichotomous
    - ordinal
    - nominal

Applicability conditions

* the observations are independent of each other (subjects are not related)
* The relationship between each predictor and the dependent variable is linear
* Residues (differences between observed and estimated values) follow a normal distribution of average 0.
* Residues have the same variability (constant variance).

Evaluation of the performance (fit to data) of the linear model:

* it is done by calculating the adjusted coefficient R2 (usually expressed as a percentage), which represents the percentage of the variability of Y that can be explained by its linear relationship with the variables X1, X2,…., Xk. The coefficient R2 is adjusted so as to allow the comparison of models with different numbers of predictors.
* Significance of the Fisher test (F-test) with the null hypothesis H0:
* Student-t tests for each partial regression coefficient H0:b1=0, H0:b2=0,…., H0:bk=0: it is relevant if we want to determine if the dependent variable depends linearly on a certain independent variable while controlling the effects of the other predictors (covariates).

**Interpretation of the coefficients of the linear model**

Consider a sample of size n and measure the value of each of the variables X1, X2,…., Xk and Y for each subject included in the sample.

The linear regression model involves testing a shape model:

Where

* is the expected / theoretical (that can be linear modelated from X1, X2,…., Xk) value of the variable of the variable Y (e.g. its average) for certain values ​​of the variables X1, X2,…., Xk and is called the estimated / predicted / adjusted value of the variable Y
* X1, X2,…., Xk are independent variables
* b0 represents the free term (model constant) and is the estimated value of the variable Y when the predictor (s) take the value 0
* b1,…, bk are partial regression coefficients which are interpreted as follows:
  + if the predictor variable (Xi) is continuous quantitative: the regression coefficient (bi) shows how many units of measure y changes (on average) when Xi changes by 1 unit after adjusting for the other predictors (or in other words, is the difference in the mean of Y comparing two subpopulations / subgroups that differ in Xi by 1 unit while keeping constant the values ​​of the other predictors in the model)
  + if the predictor variable (Xi) is dichotomous: the regression coefficient (bi) represents the difference in the mean value of the variable Y between the two groups defined by the predictor categories while keeping constant the values ​​of the other predictors in the model
  + if the predictor variable (Xi) is nominal with p categories: p-1 partial regression coefficients will be obtained, each representing the difference in the mean value of the variable Y for a certain category versus the reference category while keeping the values ​​of the other predictors in the model constant

**!!! The simplified case of evaluating the contribution of a single predictor to the variability of a dependent variable is known as univariate regression analysis. Otherwise, a multivariate regression analysis will be performed to evaluate the contribution of a set of predictors.**

## Linear regression analysis: Scenario 5

You are dissatisfied with the result of the previous article. You also want to find an article where possible confounding / confusion factors such as the age, BMI are controlled (the analysis keep constant the values of these possible factors). Look no further and find: Mahabir et al [[[4]](#footnote-4)]. Recognized BMC Publishing House recommends this article. The journal's impact factor 3,359 is quite high (Q2 in Nutrition and Dietetics, ranking 39 out of 89) and assures you that the research is a qualitative one. You expect a multivariate regression analysis that is statistically superior to a univariate regression analysis or correlation analysis.

In this study, 51 menopausal patients were selected. These were the control group in a trial with a controlled diet for 8 weeks (they ate only food provided by the research center). They were evaluated two times initially and at the end of the study. Blood sample were taken each time.

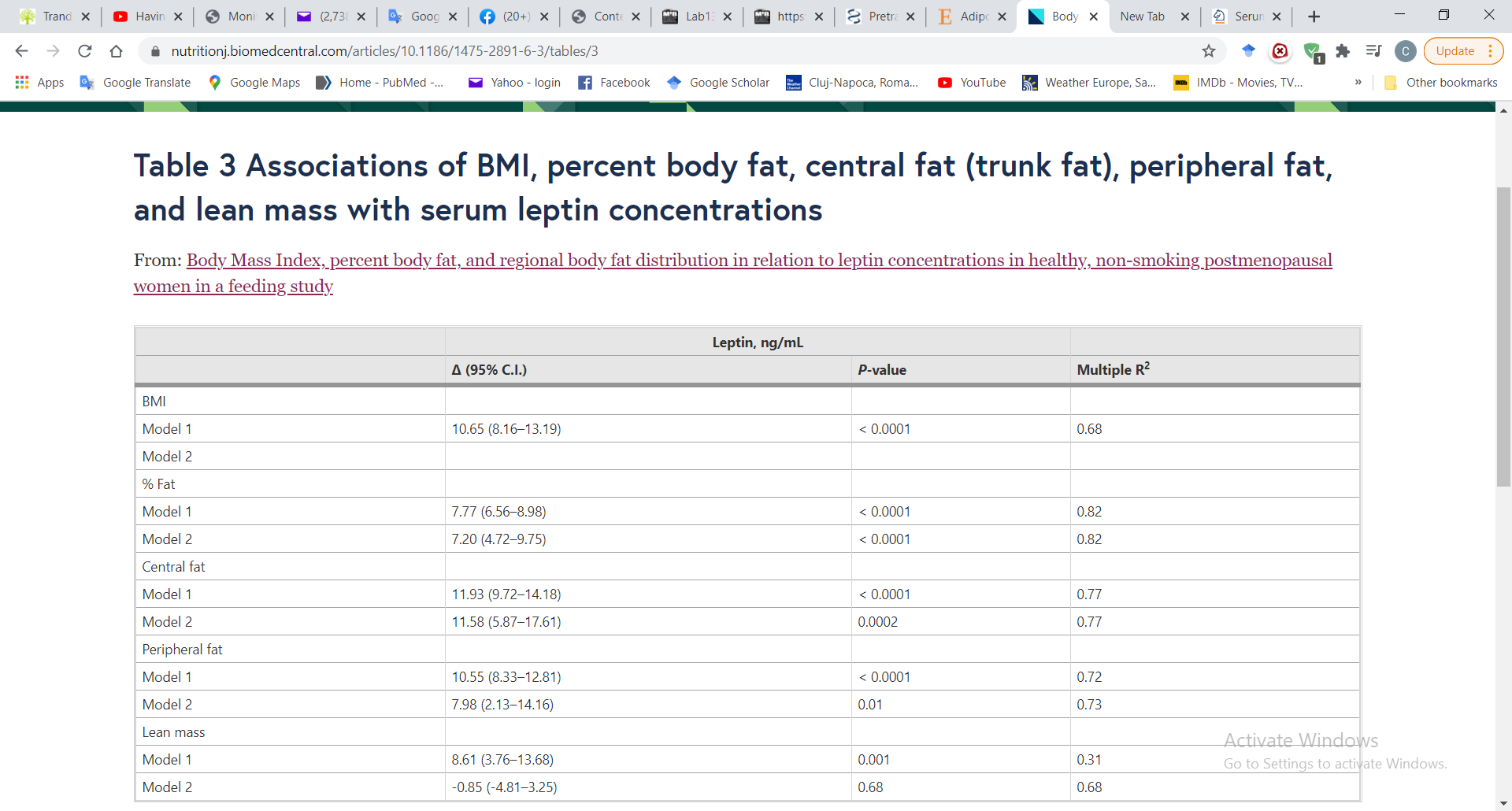
The chapter "Statistical methods" of the article mentions:

„Serum leptin concentrations were log transformed using the natural log. Linear regression models estimated percent changes in serum leptin concentrations per one-unit change in BMI, modeled as continuous variable. All models included age (continuous), parity (continuous), race (example, African American, yes / no), age of menarche (less than 12, yes / no), and family history of breast cancer (mother or full sister with breast cancer, yes / no). Multiple R2 and F-tests were calculated from the linear regression models.”[4]

**1. Specify why do you think that leptin was transformed by logarithm:**

|  |
| --- |
|  |

**2.** Find Table 3 and the associated text: “Table 3 shows how much serum leptin concentration changed for a one-unit increase in BMI. For model 1 (adjusted for age, race, family history of breast cancer, parity, and menarche <12 years), we found statistically significant increases in serum leptin concentrations for all five measures of adiposity. Leptin increased 10.7% (95% CI = 8.2% –13.2%) for each one-unit increase in BMI.” [4]



[4]

**a. Specify how R2 = 0.68 is interpreted in case of leptin dependence on BMI:**

|  |
| --- |
|  |

**b. How to interpret the 95% confidence interval: 8.2-13.2?**

|  |
| --- |
|  |

**c. Specify what in the meaning of „the model has been adjusted for by age, race, etc.” Interpret the result from this point of view:**

|  |
| --- |
|  |

**d. Specify the relationship between the coefficient 10.65 and the coefficients of the multivariate regression equation:**

|  |
| --- |
|  |

**Optional**

## Logistic regression analysis: purpose, applicability conditions, interpretation of results

**Purpose**:

* allows us to estimate the effects of several independent variables on a dichotomous type dependent variable, either to test hypotheses about predictive factors or to produce / develop / create a predictive model
* determining the probability that a subject with a certain covariate structure (with a certain combination of values ​​for the independent variables) will have a certain event of interest (e.g. the presence of a disease);

where covariate variables - any measurable independent variable that is associated with the dependent variable in a regression model.

**Data types that can be used**:

Dependent variable

• dichotomous / nominal / ordinal quality

Dependent variable

• quantitative

• qualitative dichotomous

• ordinal qualitative

• nominal quality

**Applicability conditions**

• the observations are independent of each other

• The relationship between each continuous predictor and the dependent variable is linear on the logit scale.

**Evaluation of the performance** (adequacy to data) of the logistics model:

* The insignificance of the Hosmer – Lemeshow test having as null hypothesis H0: The tested model is adequate for the data.
* The area under the ROC curve provides a measure of how well the factors included in the model predict the event of interest; it is also called the "concordance index" or "c index" (where c = 1 indicates a perfect prediction, c> 0.8 indicates a "good" predictive ability while c = 0.5 indicates a random prediction).
* Calculation of the coefficients McFadden R2, Cox-Snell R2, Nagelkerke / Cragg & Uhler’s R2

**Interpretation of the coefficients of the logistics model**

Consider a sample of size n and measure the value of each of the variables X1, X2,…, Xk and Y for each subject included in the sample.

The binomial logistics regression model uses a logarithmic transformation to model a linear relationship of the form:

where

* p represents the probability that a subject with a set of values of independent variables has the event of interest (e.g/ disease or other outcome pursued)
* X1, X2,…, Xk are independent variables
* b0 represents the free term (model constant)
* b1,…, bk are partial regression coefficients that when transformed from the logarithmic scale to the natural scale become ORs (odds ratio), so the exponential of a logistic regression coefficient is the ratio of chances (OR) to increase by 1 unit of the value of the independent variable after adjusting for the other predictors and is interpreted as follows:
  + if the predictor variable (Xi) is dichotomous:
    - , the groups being determined by the 2 categories of the predictor (if bi> 0 then >1, otherwise if bi <0 then OR <1).
  + if the predictor variable (Xi) is nominal with p categories: p-1 ORs will be obtained, each of them being the ratio between the chance of the event of interest to a certain group related to the chance of the event on the reference group; the groups being determined by the p categories of the predictor;
  + if the predictor variable (Xi) is quantitative continuous: OR represents the change in the odds of the event of interest when Xi changes by 1 unit after adjusting for the other predictors (e.g. an increase by 2 units of the predictor will be associated with an OR = OR × OR = OR2 and not 2 × OR).

## Logistic regression analysis: Scenario 6

You want to evaluate the relationship between certain factors and the onset of diabetes after a kidney transplant. You are wondering which statistical method is appropriate for your data.

You conduct an analytical study. For all subjects who underwent kidney transplantation, evaluate the characteristics of interest and evaluate the occurrence of diabetes after transplantation.

The variables of interest are: independent variables: living donor (yes / no), pre-transplant serum creatine level, age of the transplanted patient; the weight of the transplanted patient; dependent variable: presence of renal posttransplant diabetes.

You gathered a total of 134 subjects.

1. **Specify the type of data collection (control case / exposed / unexposed / representative sample)**

|  |
| --- |
|  |

1. **Specify the name and type of each study variable (dichotomous qualitative, nominal qualitative, ordinal qualitative, continuous quantitative)**

|  |
| --- |
| **living donor (yes / no),**  **pre-transplant serum creatine level (mg/dl)**  **age of the transplanted patient (years)**  **the weight of the transplanted patient (kg)**  **presence/absences of renal posttransplant diabetes** |

1. **Specify the name of the statistical method appropriate to your data:**

|  |
| --- |
|  |

## Logistic regression analysis: Scenario 7

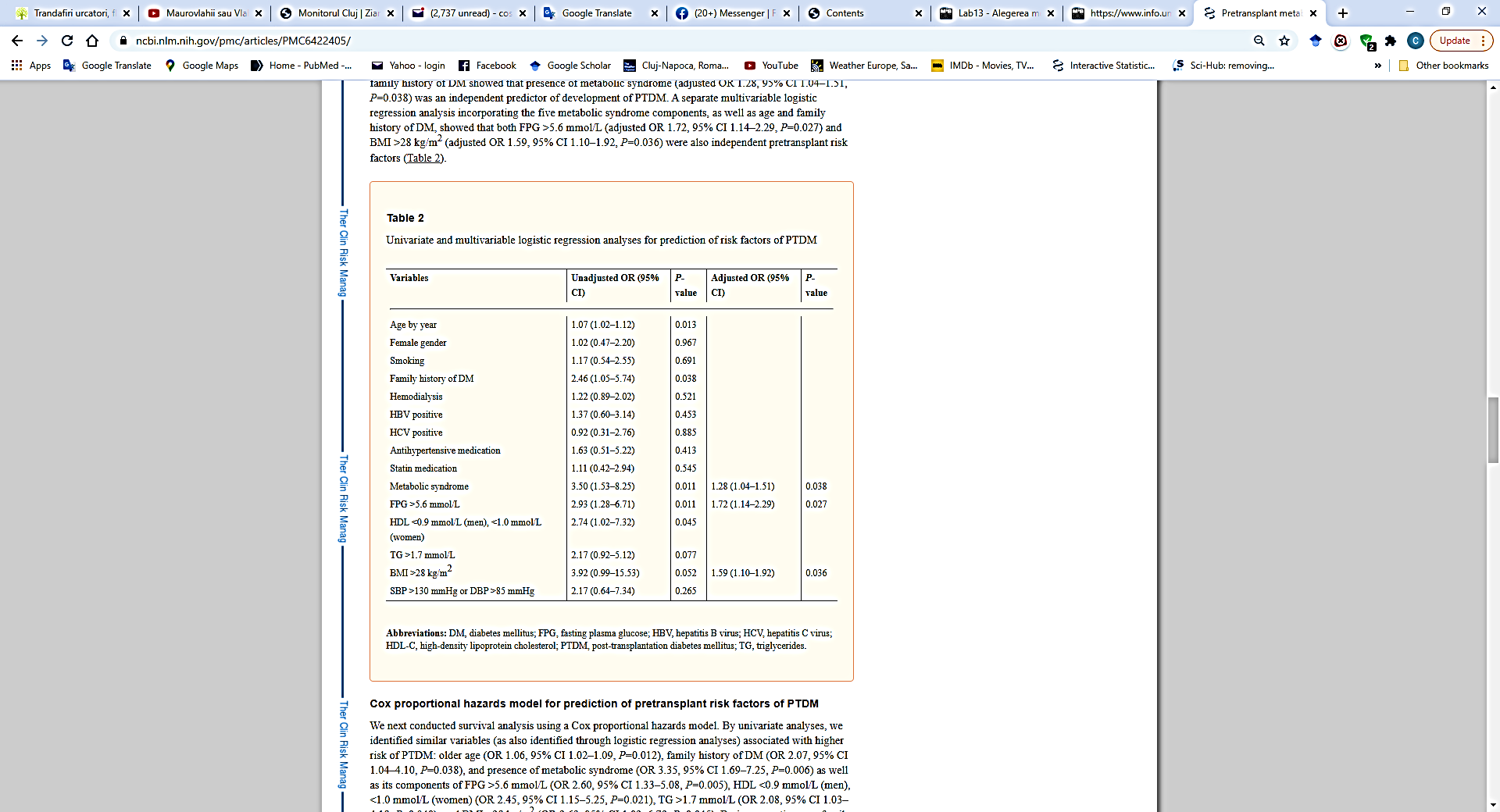
You want to evaluate the relationship between the genetic polymorphism of adiponectin and the occurrence of post-kidney diabetes (PTDM). The univariate study showed that patients with the GG genotype are more prone to PTDM than those with the T allele (GT and TT). You look in the literature for a similar study to find the factors you should control by multivariate analysis so that you can conclude that there is an association between the genetic polymorphism of adiponectin and the onset of post-kidney diabetes.

You found a published retrospective study. The study was done on a sample of 633 patients who had a kidney transplant. 26% developed PTDM within 1 year after transplantation: Cai et al [[[5]](#footnote-5)]

The authors state: “To find the potential pre-transplant risk factors for PTDM, we first performed a univariate analysis to assess the association of individual variables with the development of PTDM. Variables with a p value <0.1 were selected for further analysis using multivariate logistic regression to assess whether the relationship persisted. ”[5]

In the article you will find the following table and the text that describes it:

“To address the potential pretransplant risk factors of PTDM, we first conducted univariate logistic regression analyzes (with unadjusted OR) for all the collected variables. The analyzes showed that the following variables were associated with higher risk of PTDM (P> 0.1): older age (OR 1.07, 95% CI 1.02–1.12, P = 0.013), family history of DM (OR 2.46, 95% CI 1.05 –5.74, P = 0.038), and presence of metabolic syndrome (OR 3.50, 95% CI 1.53–8.25, P = 0.011) as well as its components of FPG> 5.6 mmol / L (OR 2.93, 95% CI 1.28–6.71 , P = 0.011), HDL <0.9 mmol / L (men), <1.0 mmol / L (women) (OR 2.74, 95% CI 1.02–7.32, P = 0.045), TG> 1.7 mmol / L (OR 2.17, 95% CI 0.92–5.12, P = 0.077), and BMI> 28 kg / m2 (OR 3.92, 95% CI 0.99–15.53, P = 0.052). Next, multivariable logistic regression analysis adjusted for age and family history of DM showed that the presence of metabolic syndrome (adjusted OR 1.28, 95% CI 1.04–1.51, P = 0.038) was an independent predictor of development of PTDM. A separate multivariable logistic regression analysis incorporating the five metabolic syndrome components, as well as age and family history of DM, showed that both FPG> 5.6 mmol / L (adjusted OR 1.72, 95% CI 1.14–2.29, P = 0.027) and BMI > 28 kg / m2 (adjusted OR 1.59, 95% CI 1.10–1.92, P = 0.036) were also independent pretransplant risk factors (Table 2). ”[5]

[5]

**1.** **Specify the univariate statistical method used to calculate the OR and the 95% confidence interval:**

|  |
| --- |
|  |

**2. Specify the multivariate statistical method used to calculate the OR and the 95% confidence interval:**

|  |
| --- |
|  |

**3. Specify the variables that significantly influence PTDM according to the univariate logistic regression analysis:**

|  |
| --- |
|  |

**4. Interpret OR = 3.50 (1.53-8.25) according to the metabolic syndrome variable:**

|  |
| --- |
|  |

**5. Specify what it means that the metabolic syndrome variable has a second OR (adjusted OR) in the column corresponding to the multivariate logistic regression analysis. Which of these should be considered:**

|  |
| --- |
|  |

**6. Interpret OR = 2.46 (1.05-5.74) according to the family history of diabetes:**

|  |
| --- |
|  |

**7. Specify what it means that the variable family history of diabetes has nothing filled in the column corresponding to the multivariate analysis:**

|  |
| --- |
|  |

**8. Interpret OR = 3.92 (0.99-15.53) according to BMI> 28kg / m2**

|  |
| --- |
|  |

**9. Interpret OR = 1.59 (1.10-1.92) according to BMI> 28kg / m2**

|  |
| --- |
|  |

**10. How do you explain that the BMI factor> 28kg / m2 was not statistically significant in the univariate analysis, but was significant in the multivariate regression analysis? Specify.**

|  |
| --- |
|  |

**11. Specify the variables that significantly influence the odds of PTDM according to the multivariate regression analysis:**

|  |
| --- |
|  |

**12. Specify which factors were included in the multivariate regression analysis to determine the adjusted OR:**

|  |
| --- |
|  |

Laboratory Conclusions: Today's paper helps you

* in the analysis of data for the thesis or for other scientific papers
* to the critical appreciation and interpretation of the studies (articles) you read

E-mail the completed Word document (after saving and closing it in advance) as an attachment. In the e-mail specify the Subject: Identification data and title of the paper.

1. Sanchez-Enriquez S, Ballesteros-Gonzalez IT, Villafán-Bernal JR, Pascoe-Gonzalez S, Rivera-Leon EA, Bastidas-Ramirez BE, et al. Serum levels of undercarboxylated osteocalcin are related to cardiovascular risk factors in patients with type 2 diabetes mellitus and healthy subjects. World J Diabetes 2017;8(1):11-7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5237813/ [↑](#footnote-ref-1)
2. Bahathiq AO. Relationship of Leptin Hormones with Body Mass Index and Waist Circumference in Saudi Female Population of the Makkah Community. Open Obes J. 2010;2:95-100. <https://benthamopen.com/contents/pdf/TOOBESJ/TOOBESJ-2-95.pdf> [↑](#footnote-ref-2)
3. Elhafeez MA, Zamzam DA, Fouad MM, Elkhawas HM, Rahman HA. Serum leptin and body mass index in a sample of Egyptian multiple sclerosis patients. The Egyptian Journal of Neurology, Psychiatry and Neurosurgery. 2020;56(107).

   <https://ejnpn.springeropen.com/articles/10.1186/s41983-020-00239-3> [↑](#footnote-ref-3)
4. Mahabir S, Baer D, Johnson LL, Roth M, Campbell W, Clevidence B et al. Body Mass Index, percent body fat, and regional body fat distribution in relation to leptin concentrations in healthy, non-smoking postmenopausal women in a feeding study. Nutr J. 2007;6(3). https://nutritionj.biomedcentral.com/articles/10.1186/1475-2891-6-3 [↑](#footnote-ref-4)
5. Cai R, Wu M, Xing Y. Pretransplant metabolic syndrome and its components predict post-transplantation diabetes mellitus in Chinese patients receiving a first renal transplant. Ther Clin Risk Manag. 2019 Mar 14;15:497-503. doi: 10.2147/TCRM.S190185. PMID: 30936711; PMCID: PMC6422405.

   <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6422405/> [↑](#footnote-ref-5)