Tests of Associations Correlations & Regressions

Sorana D. Bolboacă

OUTLINE & OBJECTIVES

OUTLINE

- Correlation methods
  - Parametric: Pearson
  - Non-parametric: Spearman, Kendall, etc.
- Regression analysis:
  - Linear methods

OBJECTIVES

- To be able to evaluate and interpret the product moment correlation coefficient and Spearman's correlation coefficient
- To be able to find and interpret the equations of regression lines
- To be able to investigate the strength and direction of a relationship between independent and dependent variables

Correlation: 3 Characteristics

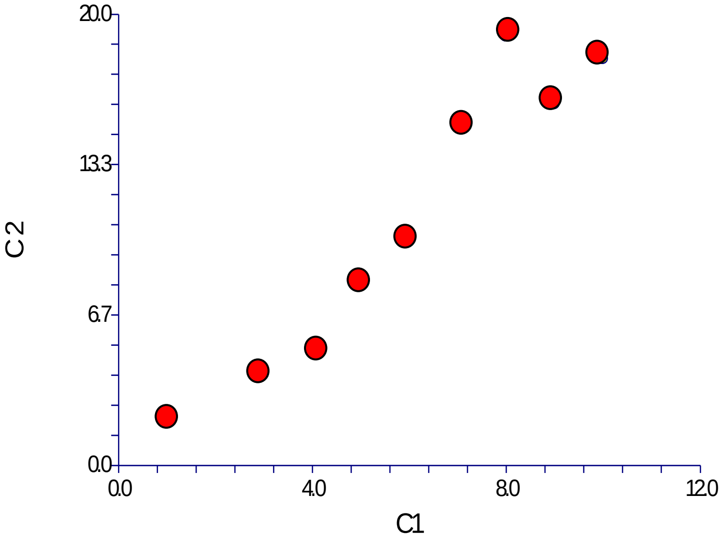
Correlation: a statistical technique that measures and describes the degree of linear relationship between two variables

1. Direction: Positive (+) vs. Negative (-)
2. Degree of association:
   - ❑ Takes values between -1 and +1
   - ❑ Absolute value = strength
3. Form: Linear vs. Non-linear

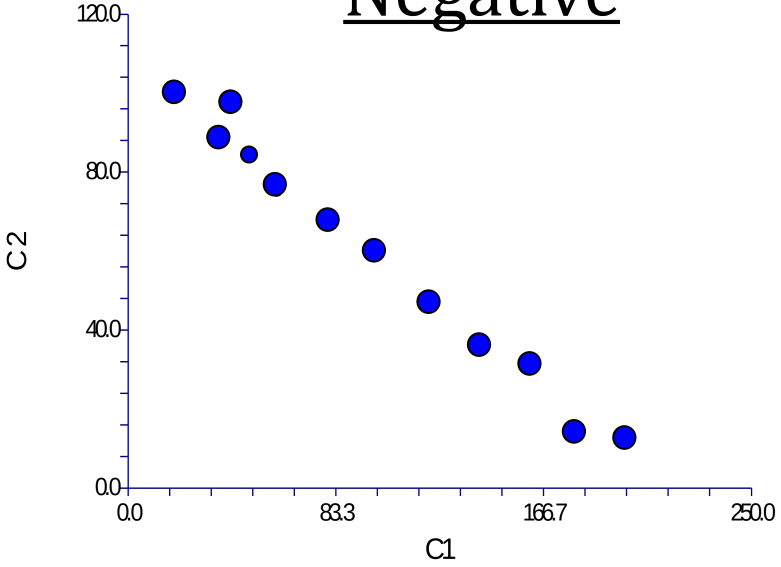
Correlation is applied on two variables

CORRELATION: 1. DIRECTION

Positive



Negative



Large values of X = large values of Y
Small values of X = small values of Y

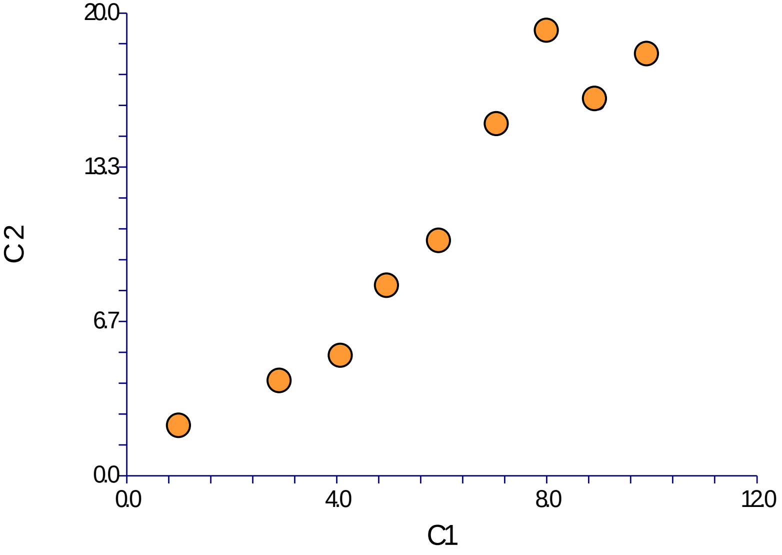
Large values of X = small values of Y
Small values of X = large values of Y

e.g. IQ (Intelligence Quotient) and SAT

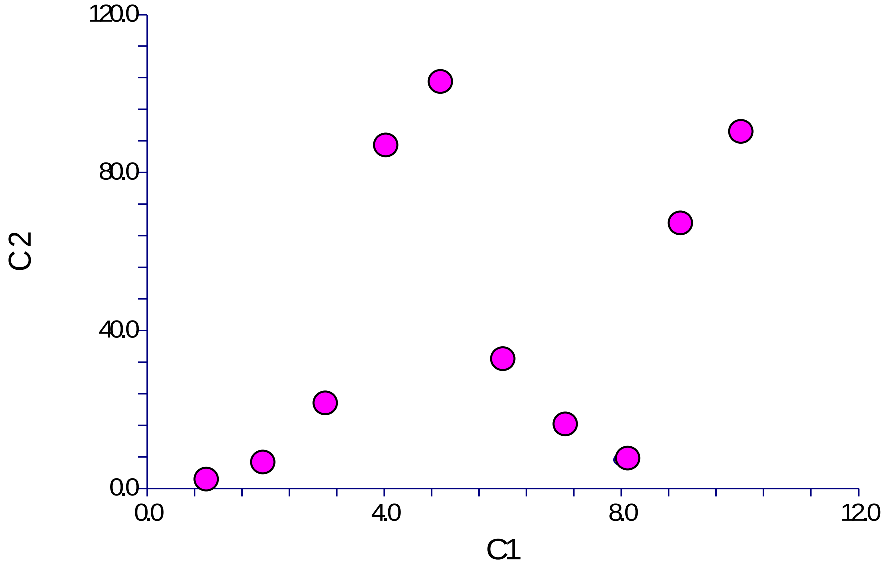
e.g. SPEED and ACCURACY

Correlation: 2. Degree of association

Strong (tight cloud)

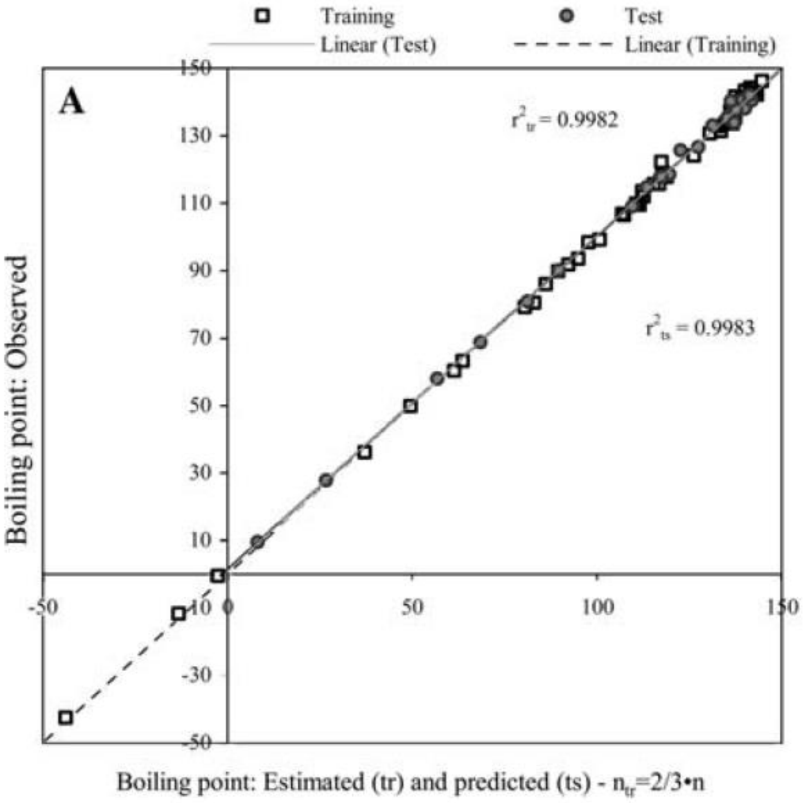


Weak (diffuse cloud)



Correlation: 3. Form

Linear



Bolboacă SD, Jäntschi L. Modelling the property of compounds from structure: statistical methods for models validation. Environmental Chemistry Letters 2008;6:175-181.

Non- linear

$\hat{y} = 0.8173 - 0.7972*\exp(-x/2.6772)$

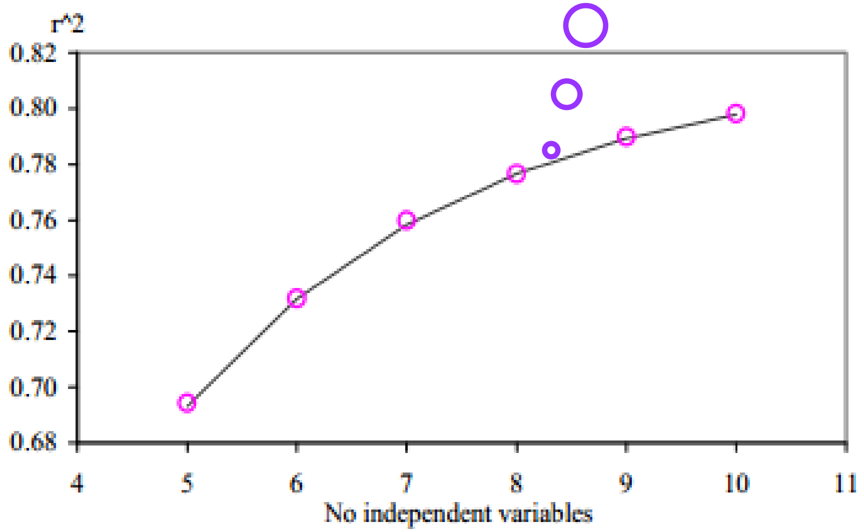


Figure 2. The dependence between r^2 and the number of independent variables for $4 < x \leq 10$

Bolboacă SD, Jäntschi L. Dependence between determination coefficient and number of regressors: a case study on retention times of mycotoxins. Studia Universitatis Babes-Bolyai Chemia 2011;LVI(1):157-166.

PEARSON CORRELATION COEFFICIENT

Symbol: r, R

A value ranging from -1.00 to 1.00 indicating the <u>strength</u> (look to the number of correlation coefficient) and <u>direction</u> (look to the sign of the correlation coefficient) of the linear relationship.

- Absolute value indicates strength
- +/- indicates direction

Sum of products

$$r = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum\left(X - \overline{X}\right)^2 \sum\left(Y - \overline{Y}\right)^2}}$$

PEARSON CORRELATION COEFFICIENT

Assumptions:

- The errors in data values are independent from one another
- Correlation always requires the assumption of a straight-line relationship
- The variables are assumed to follow a bivariate normal distribution

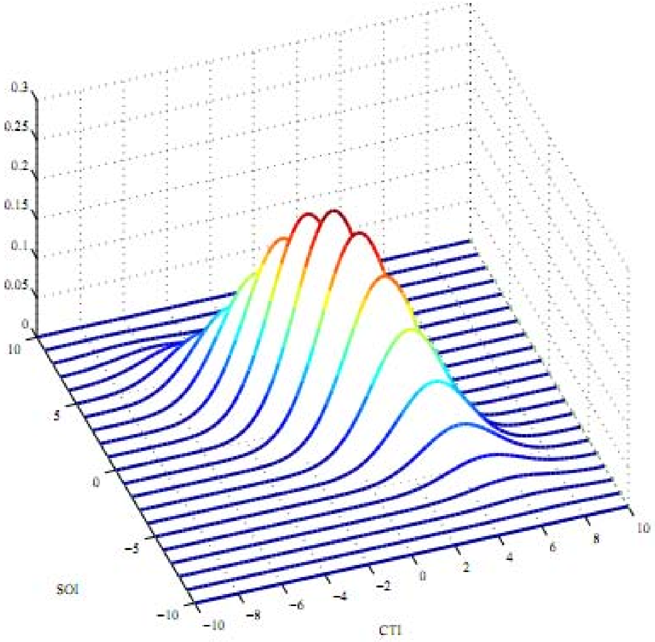
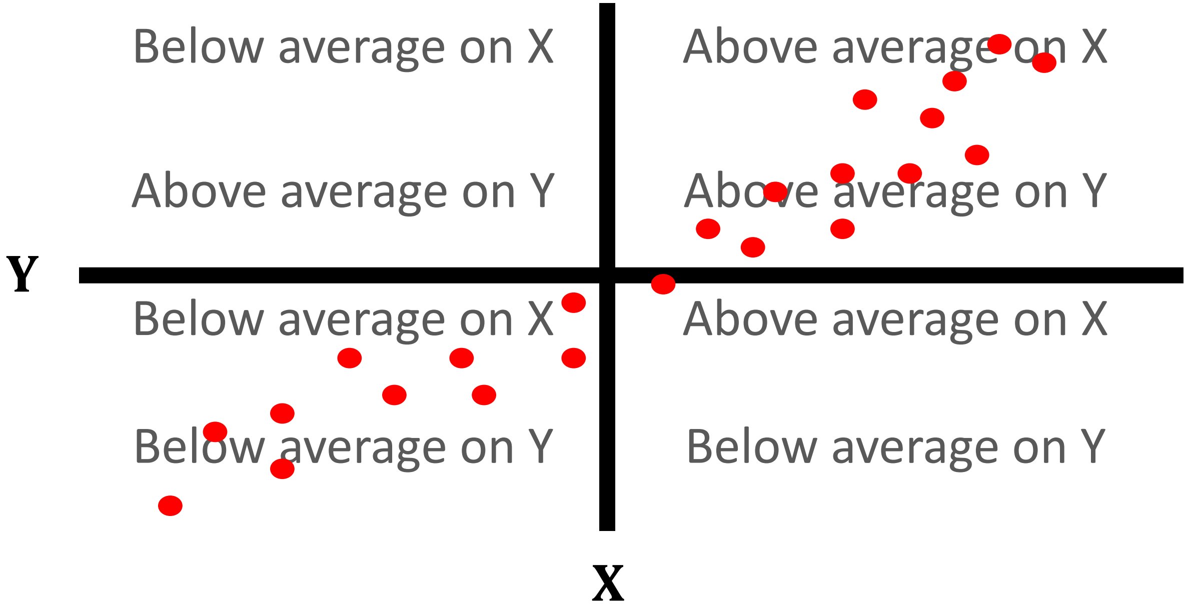


Figure 1: Bivariate Normal PDF calculated for parameters based on the Cold Tongue Index (x axis) and the Southern Oscillation Index (y-axis).

http://www.aos.wisc.edu/~dvimont/aos575/Handouts/bivariate_notes.pdf

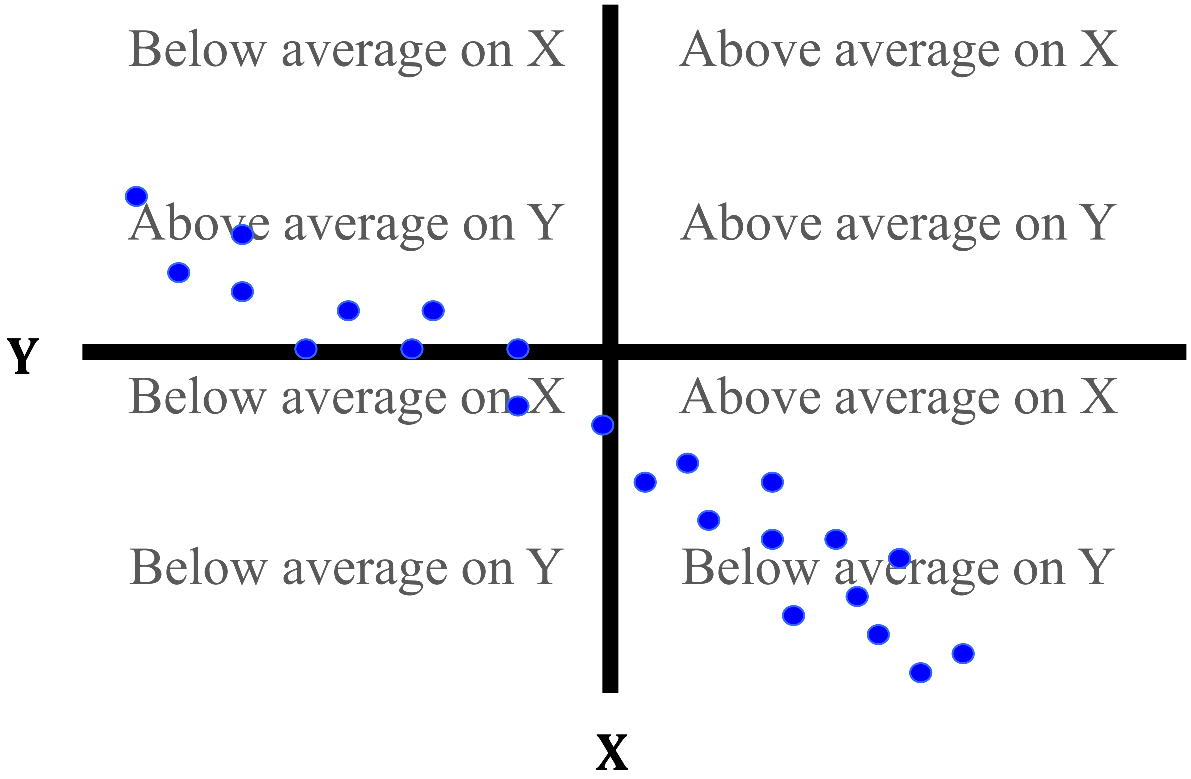
Pearson Correlation Coefficient

- For a strong <u>positive</u> association, the SP (sum of products) will be a big positive number



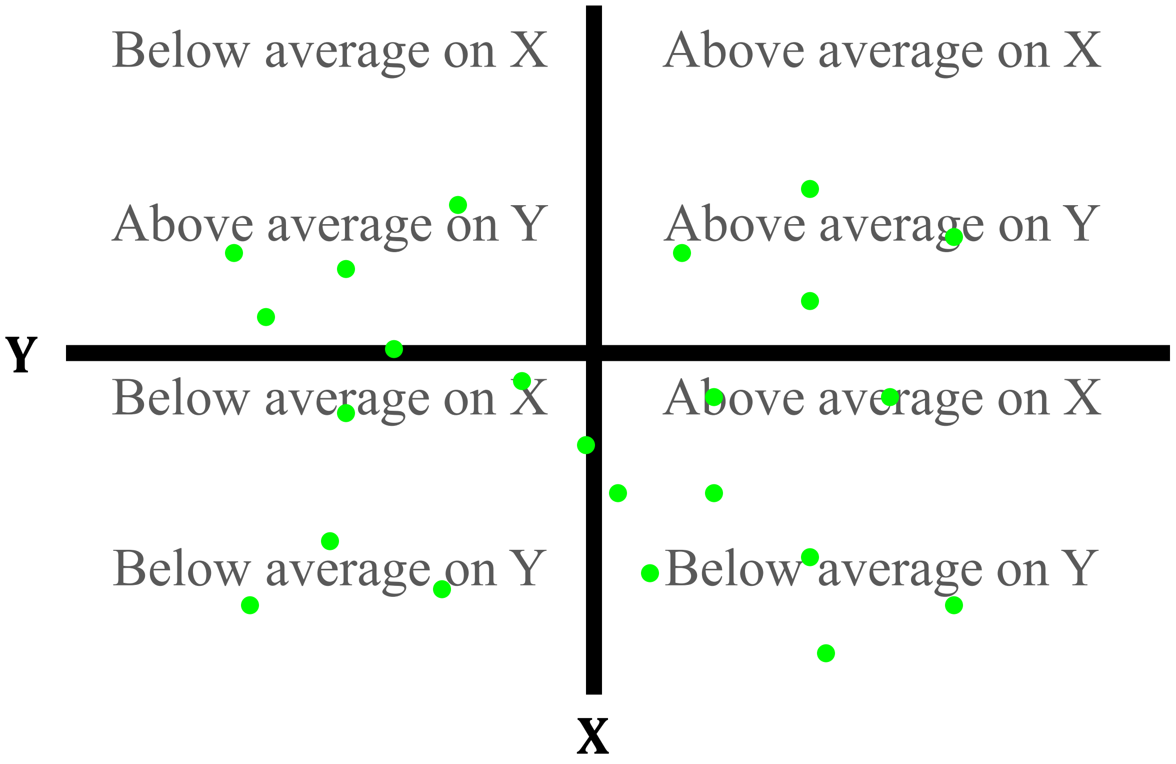
Pearson Correlation Coefficient

- For a strong negative association, the SP will be a big negative number



Pearson Correlation Coefficient

- For a weak association, the SP will be a small number (+ and – will cancel each other out)



PEARSON CORRELATION COEFFICIENT: INTERPRETATION

- A measure of strength of association: how closely do the points cluster around a line?
- A measure of the direction of association: is it positive or negative?
- Colton [Colton T. Statistics in Medicine. Little Brown and Company, New York, NY 1974] rules:
    - $R \subset [-0.25 \text{ to } +0.25] \rightarrow$ No relation
    - $R \subset (0.25 \text{ to } +0.50] \cup (-0.25 \text{ to } -0.50] \rightarrow$ weak relation
    - $R \subset (0.50 \text{ to } +0.75] \cup (-0.50 \text{ to } -0.75] \rightarrow$ moderate relation
    - $R \subset (0.75 \text{ to } +1) \cup (-0.75 \text{ to } -1) \rightarrow$ strong relation

Pearson Correlation Coefficient: Interpretation

- The P-value is the probability that you would have found the current result if the correlation coefficient were in fact zero (null hypothesis).
- If this probability is lower than the conventional significance level (e.g. 5%) ($p < 0.05$) → the correlation coefficient is called statistically significant.

- "Results: Fatigue correlated with MRCD score (Medical Research Council dyspnoea score) (r=0.57, P<0.001) and FEV(1)% predicted (r=-0.30, P=0.001)."

Hester KL, Macfarlane JG, Tedd H, Jary H, McAlinden P, Rostron L, Small T, Newton JL, De Soyza A. Fatigue in bronchiectasis. QJM. 2012;105(3):235-40.

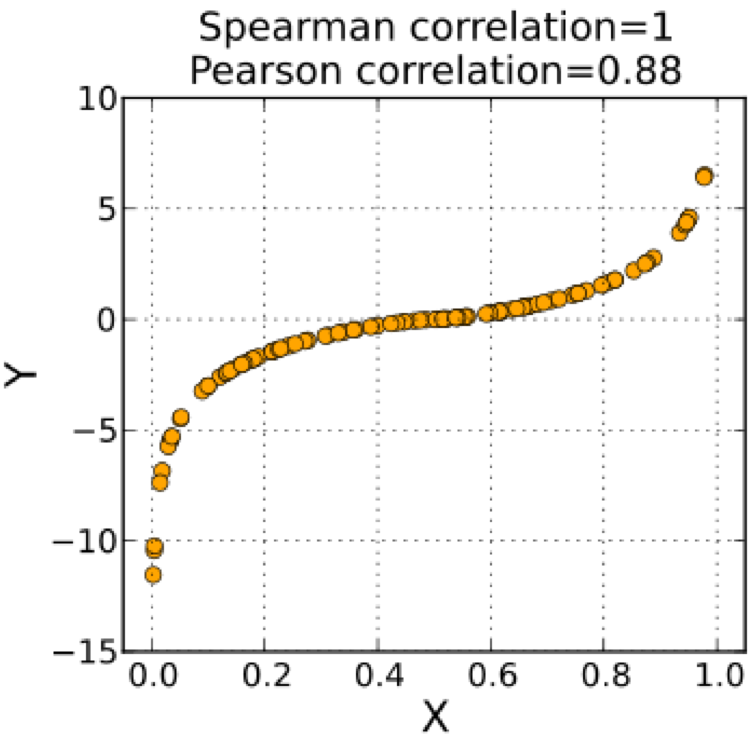
SPEARMAN RANK CORRELATION COEFFICIENT

- Not continuous measurements
- The assumption of bivariate normal distribution is violated
- Symbol: ρ (Rho Greek Letter)

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- The sign of the Spearman correlation indicates the direction of association between *X* (the independent variable) and *Y* (the dependent variable).
- ρ =1 → the two variables being compared are monotonically related. N.B. This does not give a perfect Pearson correlation.

SPEARMAN RANK CORRELATION COEFFICIENT



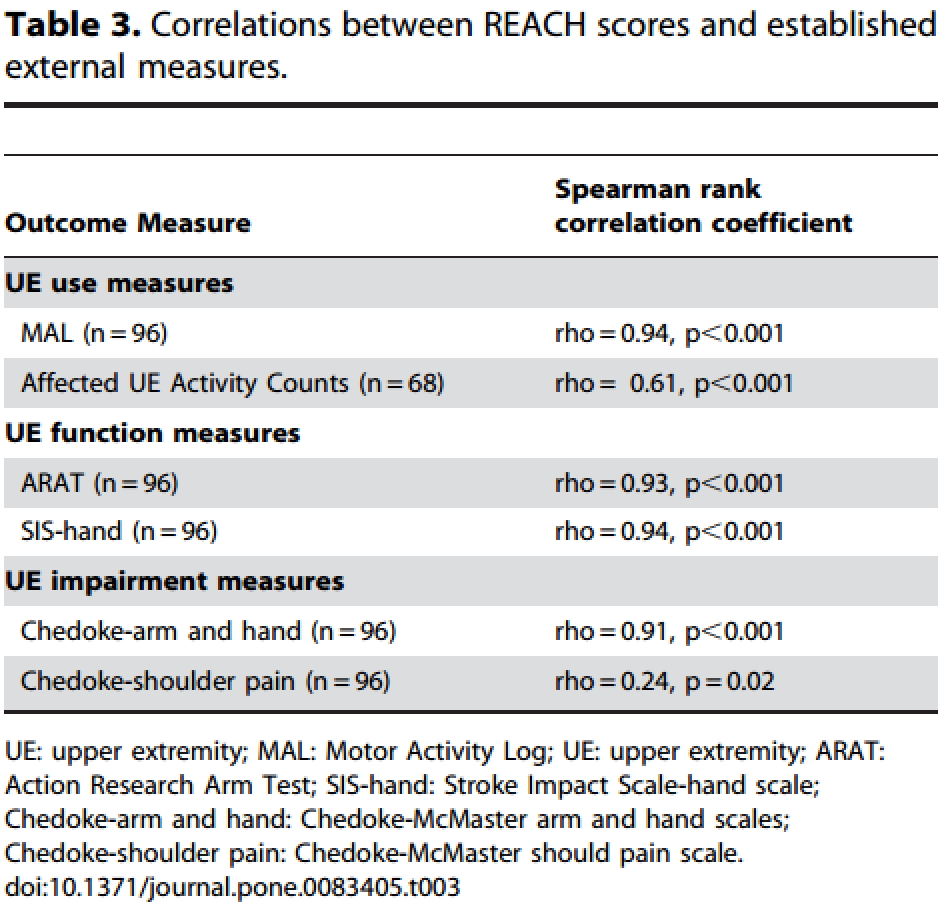


Table 3. Correlations between REACH scores and established external measures.

Outcome Measure	Spearman rank correlation coefficient
UE use measures	
MAL (n = 96)	rho = 0.94, p<0.001
Affected UE Activity Counts (n = 68)	rho = 0.61, p<0.001
UE function measures	
ARAT (n = 96)	rho = 0.93, p<0.001
SIS-hand (n = 96)	rho = 0.94, p<0.001
UE impairment measures	
Chedoke-arm and hand (n = 96)	rho = 0.91, p<0.001
Chedoke-shoulder pain (n = 96)	rho = 0.24, p = 0.02

UE: upper extremity; MAL: Motor Activity Log; UE: upper extremity; ARAT: Action Research Arm Test; SIS-hand: Stroke Impact Scale-hand scale; Chedoke-arm and hand: Chedoke-McMaster arm and hand scales; Chedoke-shoulder pain: Chedoke-McMaster should pain scale.
doi:10.1371/journal.pone.0083405.t003

PROPERTIES OF CORRELATION COEFFICIENT

- A standardized statistic – will not change if you change the units of X or Y.
- The same whether X is correlated with Y or vice versa
- Fairly unstable with small n
- Vulnerable to outliers
- Has a skewed distribution

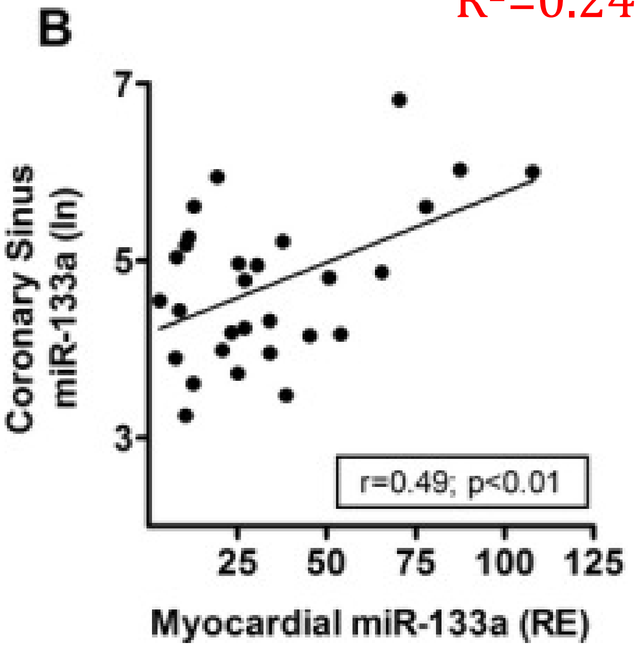
Interpretation of R-squared (R^2)

- The amount of covariation compared to the amount of total variation.

  $$R^2 = \text{explained variance} / \text{overall variance}$$

- The percent of total variance that is shared variance.
- E.g. If r = 0.80, then X explains 64% of the variability in Y (and vice versa)

$R^2=0.24$



García R, Villar AV, Cobo M, Llano M, Martín-Durán R, Hurlé MA, Francisco Nistal J. Circulating levels of miR-133a predict the regression potential of left ventricular hypertrophy after valve replacement surgery in patients with aortic stenosis. J Am Heart Assoc. 2013;2(4):e000211.

REGRESSION ANALYSIS

- Multiple linear regression (normally distributed outcome)
- Logistic regression (binary outcomes)
- Cox proportional hazards regression (the outcome is time-to-event)

MULTIVARIATE REGRESSION MODELS BY EXAMPLE

Outcome	Example	Regression	Eq.	Significance of coefficients
Continuous	Blood pressure	Linear	BP(mmHg)= α + βage(years) + βsalt(tps/day)+ βsmoker(yes/no)	***slopes*** tells how much the outcome variable increases for every 1-unit increase in each predictor
Binary	High blood pressure (yes/no)	Logistic	ln (odds of high blood pressure) = α + βage(years) + βsalt(tps/day)+ βsmoker(yes/no)	***odds ratio*** tells how much the odds of the outcome increase for every 1-unit increase in each predictor
Time-to-event	Time-to-stoke	Cox	ln (rate of stoke) = α + βage(years) + βsalt(tps/day)+ βsmoker(yes/no)	***hazard ratio*** tells how much the rate of the outcome increases for every 1-unit increase in each predictor

REGRESSION ANALYSIS

- Many (independent) variables – Which to be selected in the model?
- Different outcome variable (continuous, binary, time-related)

- Important: 5 to 20 variable (at least 10 subject for variable) & *n* & “sufficient”
- Aims:
  - Identification of important predictors (independent variables) – the number of independent variables should be as smallest as possible
  - Prediction of the outcome of interest
  - Stratification by risk
  - …

LINEAR REGRESSION

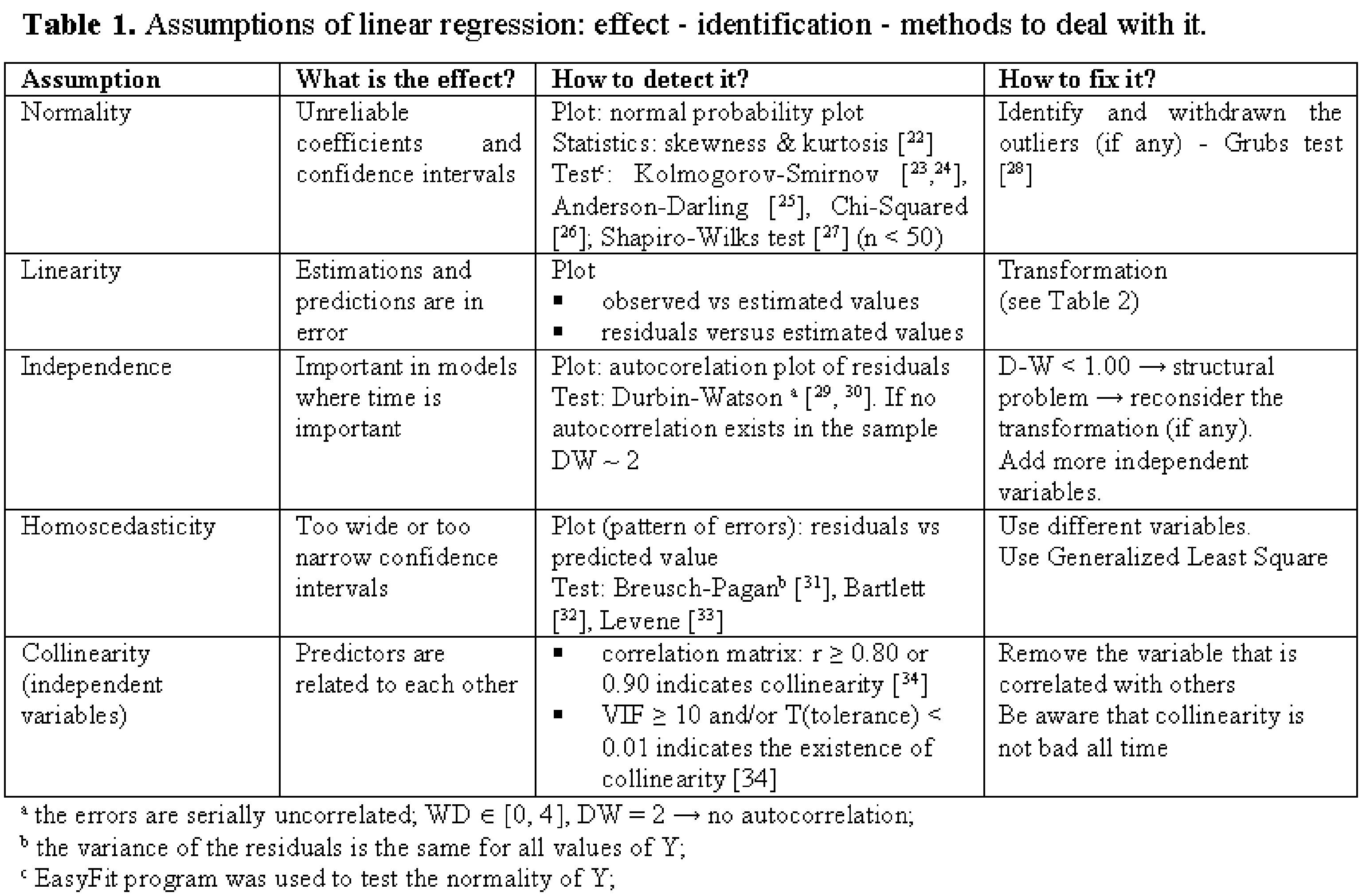


Table 1. Assumptions of linear regression: effect - identification - methods to deal with it.

Assumption	What is the effect?	How to detect it?	How to fix it?
Normality	Unreliable coefficients and confidence intervals	Plot: normal probability plot Statistics: skewness & kurtosis [22] Test[c]: Kolmogorov-Smirnov [23,24], Anderson-Darling [25], Chi-Squared [26]; Shapiro-Wilks test [27] ($n < 50$)	Identify and withdrawn the outliers (if any) - Grubs test [28]
Linearity	Estimations and predictions are in error	Plot ▪ observed vs estimated values ▪ residuals versus estimated values	Transformation (see Table 2)
Independence	Important in models where time is important	Plot: autocorelation plot of residuals Test: Durbin-Watson [a] [29, 30]. If no autocorrelation exists in the sample DW ~ 2	D-W < 1.00 → structural problem → reconsider the transformation (if any). Add more independent variables.
Homoscedasticity	Too wide or too narrow confidence intervals	Plot (pattern of errors): residuals vs predicted value Test: Breusch-Pagan[b] [31], Bartlett [32], Levene [33]	Use different variables. Use Generalized Least Square
Collinearity (independent variables)	Predictors are related to each other	▪ correlation matrix: $r \geq 0.80$ or 0.90 indicates collinearity [34] ▪ VIF ≥ 10 and/or T(tolerance) < 0.01 indicates the existence of collinearity [34]	Remove the variable that is correlated with others Be aware that collinearity is not bad all time

[a] the errors are serially uncorrelated; WD ∈ [0, 4], DW = 2 → no autocorrelation;

[b] the variance of the residuals is the same for all values of Y;

[c] EasyFit program was used to test the normality of Y;

Bolboacă SD, Jäntschi L. Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example. International Journal on Mathematical Methods and Models in Biosciences (BIOMATH) 2013;2(1):1309089.

LINEAR REGRESSION

Unusual data: not identify by usual parameter (r, F)

- Outlier:
  - X's or Y
  - Regression outlier: ↑ |residuals|
- Leverage point: unusual combination of variables

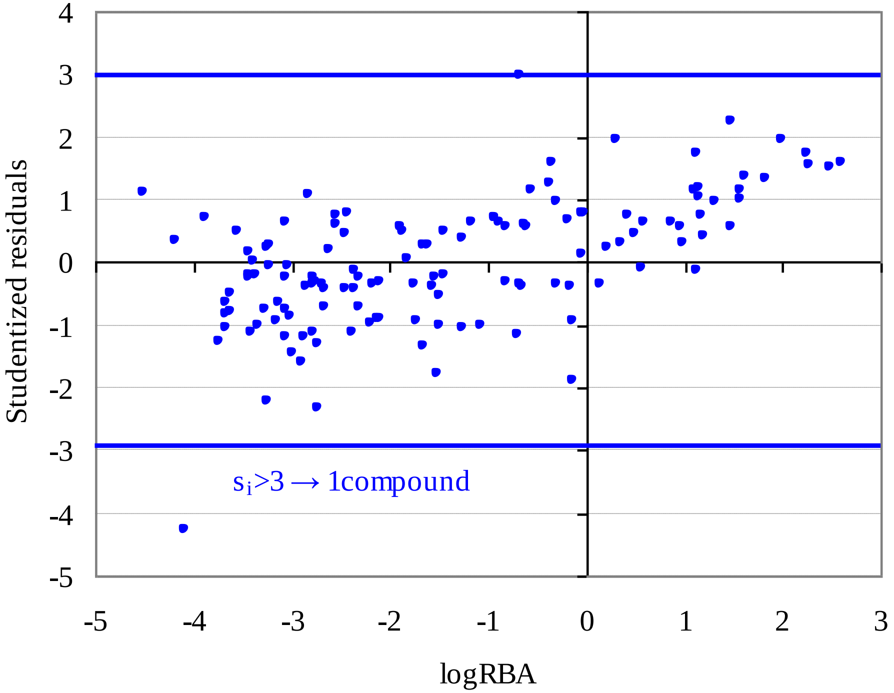
  $h_i = \boldsymbol{2 \cdot (k+1)/n}$
- Influential point: influence on the regression coefficients

  D_i model – threshold = ***4/n***

- Neither ignore, nor throw them without thinking
- Think of reason why observation may be different
- Change the model
- Fit the model with and without the unusual data and see the effect

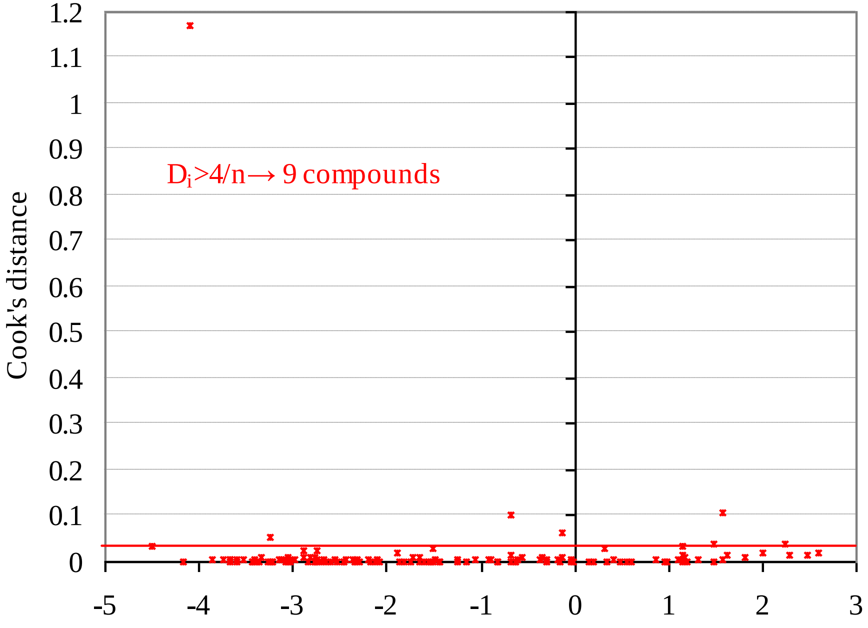
LINEAR REGRESSION

Studentized residuals



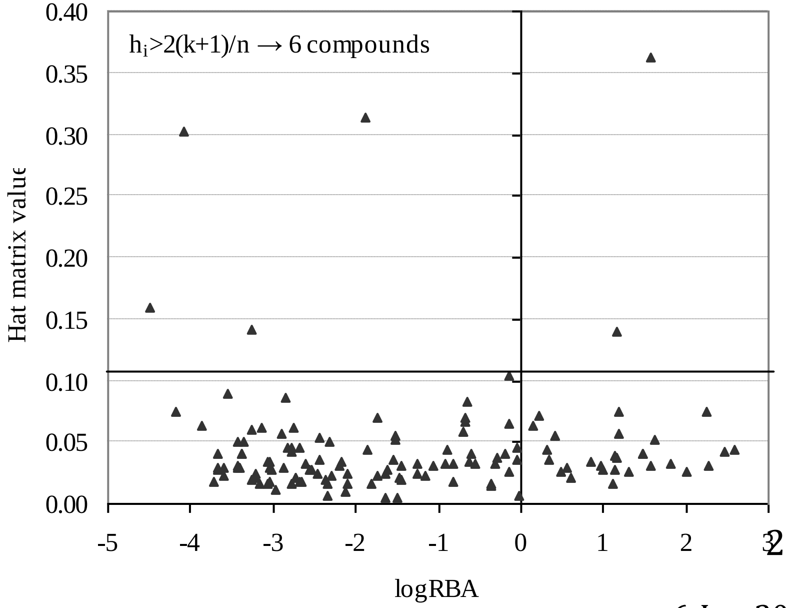
$s_i > 3 \rightarrow$ 1compound

Cook's distance



$D_i > 4/n \rightarrow$ 9 compounds

Hat matrix



$h_i > 2(k+1)/n \rightarrow$ 6 compounds

LINEAR REGRESSION DIAGNOSIS

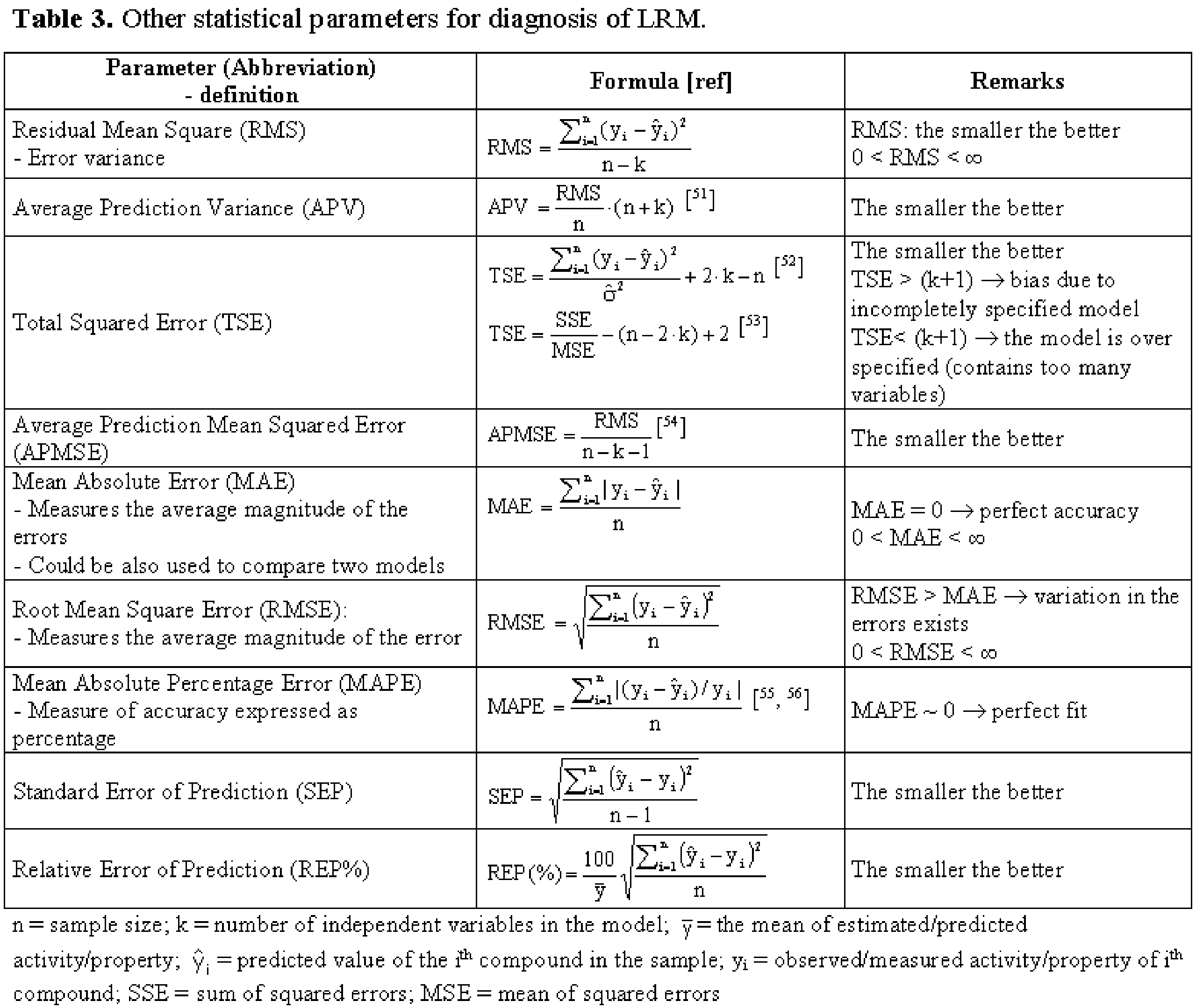


Table 3. Other statistical parameters for diagnosis of LRM.

Parameter (Abbreviation) - definition	**Formula [ref]**	**Remarks**
Residual Mean Square (RMS) - Error variance	$\text{RMS} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}$	RMS: the smaller the better $0 < \text{RMS} < \infty$
Average Prediction Variance (APV)	$\text{APV} = \frac{\text{RMS}}{n}\cdot(n+k)$ [51]	The smaller the better
Total Squared Error (TSE)	$\text{TSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\hat{\sigma}^2} + 2\cdot k - n$ [52] $\text{TSE} = \frac{\text{SSE}}{\text{MSE}} - (n - 2\cdot k) + 2$ [53]	The smaller the better TSE > (k+1) → bias due to incompletely specified model TSE< (k+1) → the model is over specified (contains too many variables)
Average Prediction Mean Squared Error (APMSE)	$\text{APMSE} = \frac{\text{RMS}}{n-k-1}$ [54]	The smaller the better
Mean Absolute Error (MAE) - Measures the average magnitude of the errors - Could be also used to compare two models	$\text{MAE} = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{n}$	MAE = 0 → perfect accuracy $0 < \text{MAE} < \infty$
Root Mean Square Error (RMSE): - Measures the average magnitude of the error	$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	RMSE > MAE → variation in the errors exists $0 < \text{RMSE} < \infty$
Mean Absolute Percentage Error (MAPE) - Measure of accuracy expressed as percentage	$\text{MAPE} = \frac{\sum_{i=1}^{n}\lvert(y_i - \hat{y}_i)/y_i\rvert}{n}$ [55, 56]	MAPE ~ 0 → perfect fit
Standard Error of Prediction (SEP)	$\text{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-1}}$	The smaller the better
Relative Error of Prediction (REP%)	$\text{REP}(\%) = \frac{100}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$	The smaller the better

n = sample size; k = number of independent variables in the model; $\bar{y}$ = the mean of estimated/predicted activity/property; $\hat{y}_i$ = predicted value of the ith compound in the sample; y_i = observed/measured activity/property of ith compound; SSE = sum of squared errors; MSE = mean of squared errors

Bolboacă SD, Jäntschi L. Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example. International Journal on Mathematical Methods and Models in Biosciences (BIOMATH) 2013;2(1):1309089.

Linear Regression Model by Example

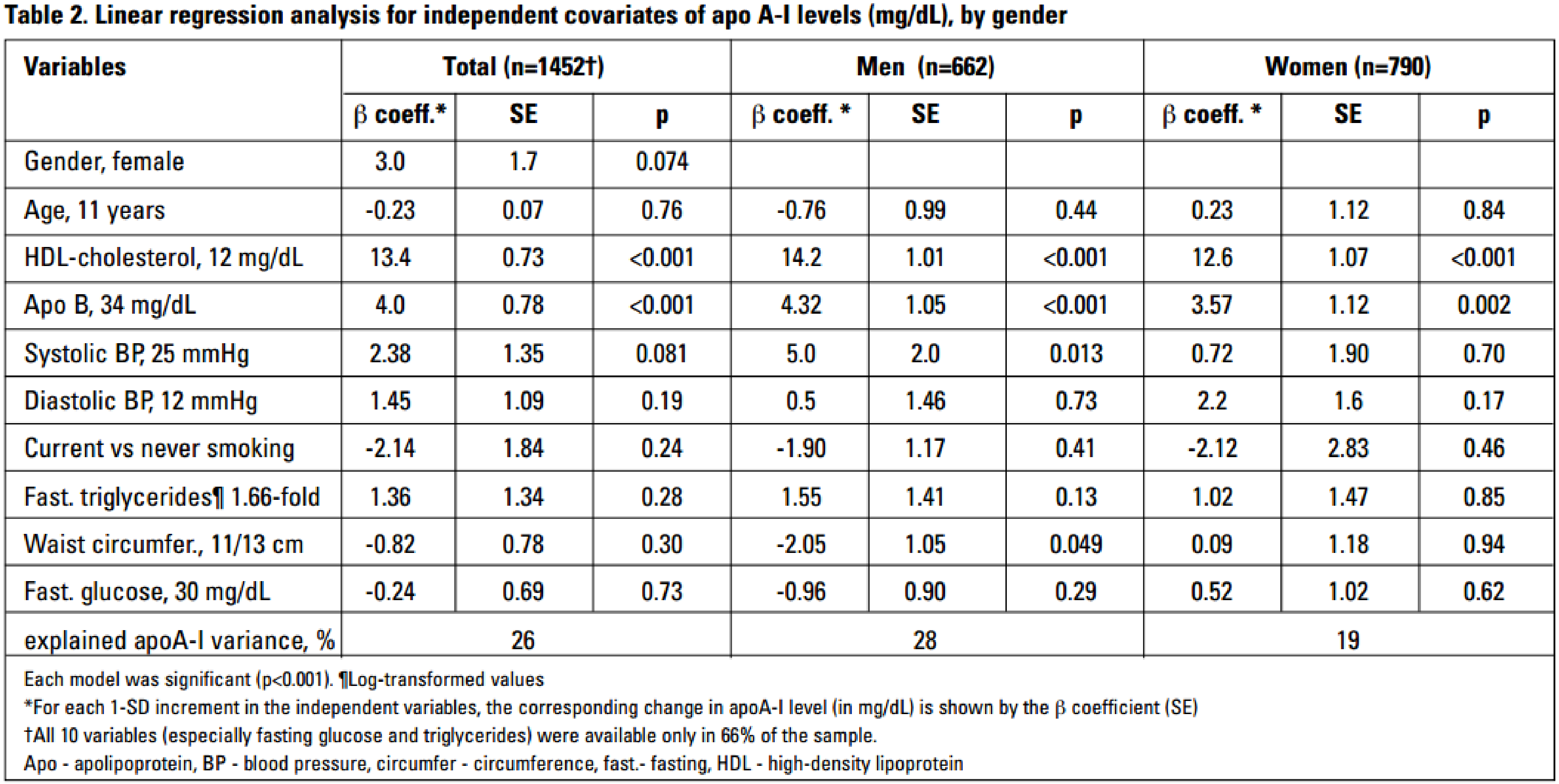


Table 2. Linear regression analysis for independent covariates of apo A-I levels (mg/dL), by gender

Variables	Total (n=1452†)			Men (n=662)			Women (n=790)		
	β coeff.*	SE	p	β coeff. *	SE	p	β coeff. *	SE	p
Gender, female	3.0	1.7	0.074						
Age, 11 years	-0.23	0.07	0.76	-0.76	0.99	0.44	0.23	1.12	0.84
HDL-cholesterol, 12 mg/dL	13.4	0.73	<0.001	14.2	1.01	<0.001	12.6	1.07	<0.001
Apo B, 34 mg/dL	4.0	0.78	<0.001	4.32	1.05	<0.001	3.57	1.12	0.002
Systolic BP, 25 mmHg	2.38	1.35	0.081	5.0	2.0	0.013	0.72	1.90	0.70
Diastolic BP, 12 mmHg	1.45	1.09	0.19	0.5	1.46	0.73	2.2	1.6	0.17
Current vs never smoking	-2.14	1.84	0.24	-1.90	1.17	0.41	-2.12	2.83	0.46
Fast. triglycerides¶ 1.66-fold	1.36	1.34	0.28	1.55	1.41	0.13	1.02	1.47	0.85
Waist circumfer., 11/13 cm	-0.82	0.78	0.30	-2.05	1.05	0.049	0.09	1.18	0.94
Fast. glucose, 30 mg/dL	-0.24	0.69	0.73	-0.96	0.90	0.29	0.52	1.02	0.62
explained apoA-I variance, %	26			28			19		

Each model was significant (p<0.001). ¶Log-transformed values

*For each 1-SD increment in the independent variables, the corresponding change in apoA-I level (in mg/dL) is shown by the β coefficient (SE)

†All 10 variables (especially fasting glucose and triglycerides) were available only in 66% of the sample.

Apo - apolipoprotein, BP - blood pressure, circumfer - circumference, fast.- fasting, HDL - high-density lipoprotein

Onat A, Can G, Örnek E, Çiçek G, Murat SN, Yüksel H. Increased apolipoprotein A-I levels mediate the development of prehypertension among Turks. Anadolu Kardiyol Derg. 2013;13(4):306-14.

LOGISTIC REGRESSION MODEL BY EXAMPLE

IF 95%CI did not contain the value of 1, the variable is a risk factor for the outcome

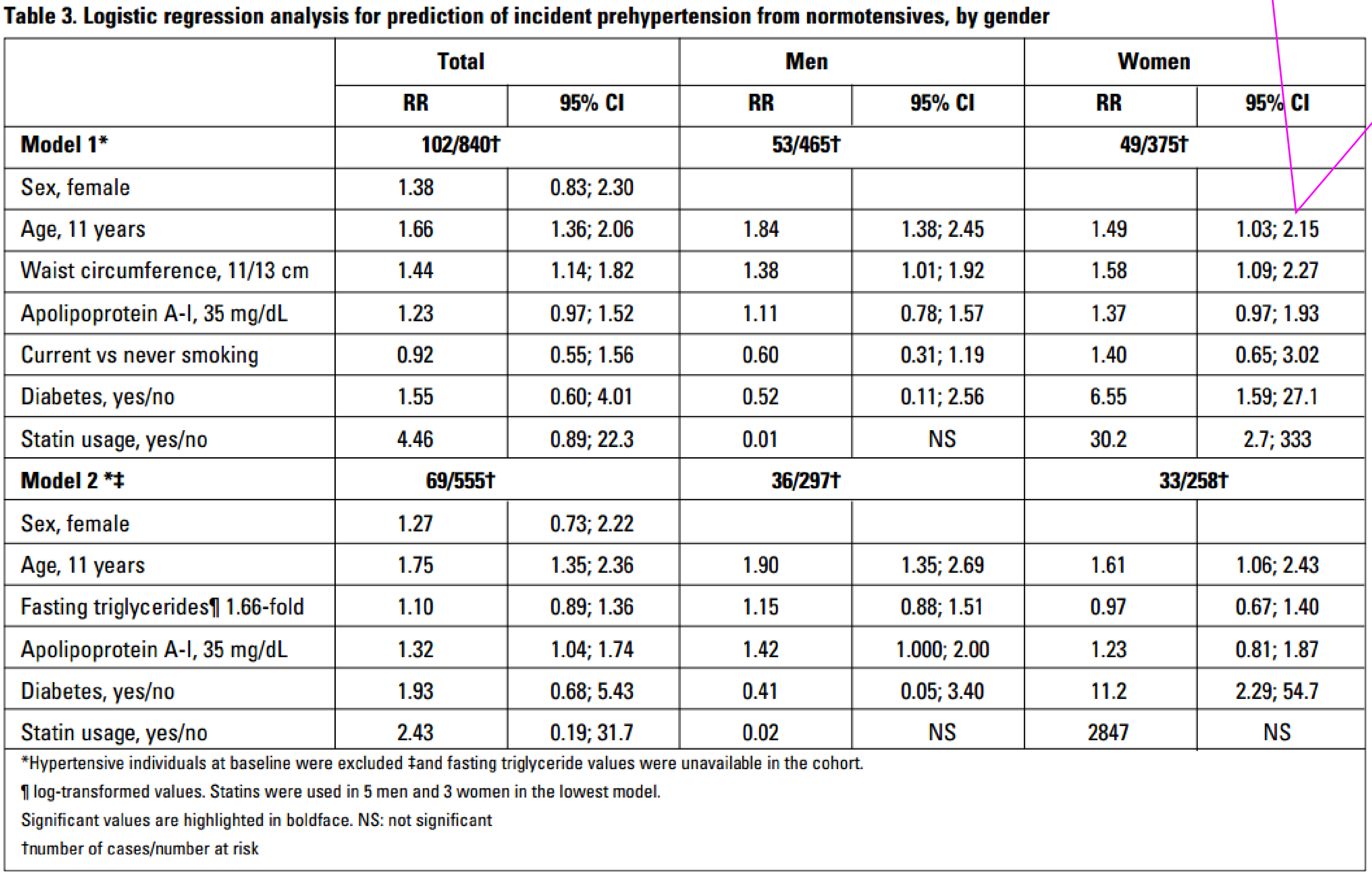


Table 3. Logistic regression analysis for prediction of incident prehypertension from normotensives, by gender

	Total		Men		Women	
	RR	**95% CI**	**RR**	**95% CI**	**RR**	**95% CI**
Model 1*	**102/840†**		**53/465†**		**49/375†**	
Sex, female	1.38	0.83; 2.30				
Age, 11 years	1.66	1.36; 2.06	1.84	1.38; 2.45	1.49	1.03; 2.15
Waist circumference, 11/13 cm	1.44	1.14; 1.82	1.38	1.01; 1.92	1.58	1.09; 2.27
Apolipoprotein A-I, 35 mg/dL	1.23	0.97; 1.52	1.11	0.78; 1.57	1.37	0.97; 1.93
Current vs never smoking	0.92	0.55; 1.56	0.60	0.31; 1.19	1.40	0.65; 3.02
Diabetes, yes/no	1.55	0.60; 4.01	0.52	0.11; 2.56	6.55	1.59; 27.1
Statin usage, yes/no	4.46	0.89; 22.3	0.01	NS	30.2	2.7; 333
Model 2 *‡	**69/555†**		**36/297†**		**33/258†**	
Sex, female	1.27	0.73; 2.22				
Age, 11 years	1.75	1.35; 2.36	1.90	1.35; 2.69	1.61	1.06; 2.43
Fasting triglycerides¶ 1.66-fold	1.10	0.89; 1.36	1.15	0.88; 1.51	0.97	0.67; 1.40
Apolipoprotein A-I, 35 mg/dL	1.32	1.04; 1.74	1.42	1.000; 2.00	1.23	0.81; 1.87
Diabetes, yes/no	1.93	0.68; 5.43	0.41	0.05; 3.40	11.2	2.29; 54.7
Statin usage, yes/no	2.43	0.19; 31.7	0.02	NS	2847	NS

*Hypertensive individuals at baseline were excluded ‡and fasting triglyceride values were unavailable in the cohort.

¶ log-transformed values. Statins were used in 5 men and 3 women in the lowest model.

Significant values are highlighted in boldface. NS: not significant

†number of cases/number at risk

Onat A, Can G, Örnek E, Çiçek G, Murat SN, Yüksel H. Increased apolipoprotein A-I levels mediate the development of prehypertension among Turks. Anadolu Kardiyol Derg. 2013;13(4):306-14.

COX REGRESSION

Statistically significant hazard ratios (HR) did not include the value of 1 in their confidence intervals

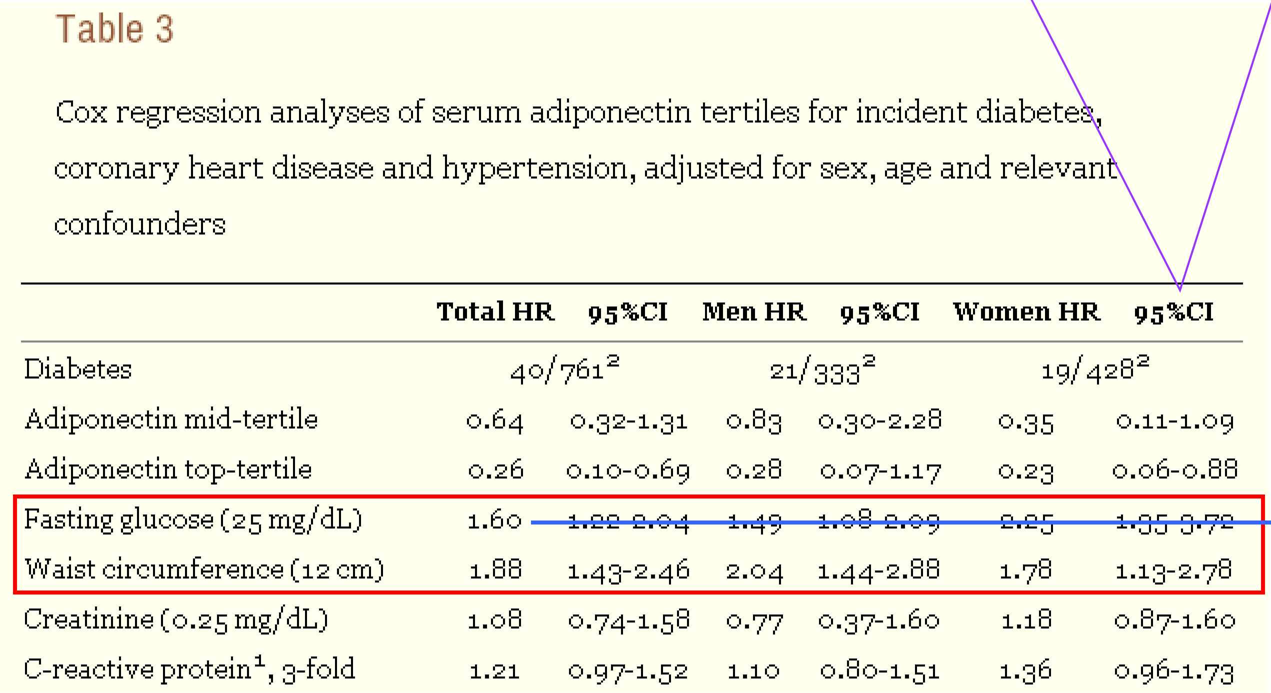


Table 3

Cox regression analyses of serum adiponectin tertiles for incident diabetes, coronary heart disease and hypertension, adjusted for sex, age and relevant confounders

	Total HR	95%CI	Men HR	95%CI	Women HR	95%CI
Diabetes	40/761[2]		21/333[2]		19/428[2]	
Adiponectin mid-tertile	0.64	0.32-1.31	0.83	0.30-2.28	0.35	0.11-1.09
Adiponectin top-tertile	0.26	0.10-0.69	0.28	0.07-1.17	0.23	0.06-0.88
Fasting glucose (25 mg/dL)	1.60	1.22-2.04	1.49	1.08-2.09	2.25	1.35-3.72
Waist circumference (12 cm)	1.88	1.43-2.46	2.04	1.44-2.88	1.78	1.13-2.78
Creatinine (0.25 mg/dL)	1.08	0.74-1.58	0.77	0.37-1.60	1.18	0.87-1.60
C-reactive protein[1], 3-fold	1.21	0.97-1.52	1.10	0.80-1.51	1.36	0.96-1.73

Group 1 (adiponectin tertiles > threshold) has a 60% higher hazard than the reference group

Onat A, Aydın M, Can G, Köroğlu B, Karagöz A, Altay S. High adiponectin levels fail to protect against the risk of hypertension and, in women, against coronary disease: involvement in autoimmunity? World J Diabetes. 2013;4(5):219-25.

INFERENTIAL STATISTICS: SUMMARY

CONTINUOUS OUTCOME VARIABLE

Are the observations independent or correlated?		Alternatives if normality is violated (± small n):
independent	correlated	
T-test: compares means between two independent groups	**Paired t-test:** compares means in paired samples	Non-parametric statistics **Wilcoxon sign-rank test**: non-parametric alternative to the paired t-test
ANOVA: compares means between > 2 independent groups	**Repeated-measures ANOVA:** compares changes over time in the means of two or more groups (repeated measurements)	**Wilcoxon sum-rank test** (=Mann-Whitney test): non-parametric alternative to the t-test
Pearson's correlation coefficient: shows linear correlation between two continuous variables	**Mixed models/GEE modeling**: multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	**Kruskal-Wallis test:** non-parametric alternative to ANOVA
Linear regression: univariate / multivariate regression technique used when the outcome is continuous; gives slopes		**Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient

BINARY (top) / TIME-TO-EVENT (bottom) OUTCOME VARIABLE

Are the observations independent or correlated?		Alternatives if normality is violated (± small n):
independent	correlated	
Chi-square test: compares proportions between two or more groups **Relative risks:** odds ratio or risk ratio **Logistic regression:** multivariate-adjusted odds ratios	**McNemar's Chi-square test:** compares binary outcome between paired groups **Conditional logistic regression** matched data **GEE modeling:** multivariate regression technique for a binary outcome when repeated measures exists	**Fisher's exact test:** compares proportions between independent groups when there are sparse data (some cells <5). **McNemar's exact test:** compares proportions between correlated groups when there are sparse data (some cells <5).

Are the observations independent or correlated?		Alternatives if normality is violated (± small n):
independent	correlated	
Kaplan-Meier statistics: estimates survival functions for each group & compares survival functions with log-rank test **Cox regression:** gives multivariate-adjusted hazard ratios	na	Time-dependent predictors or time-dependent hazard ratios (tricky!)