Correlation and Regression Analysis

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

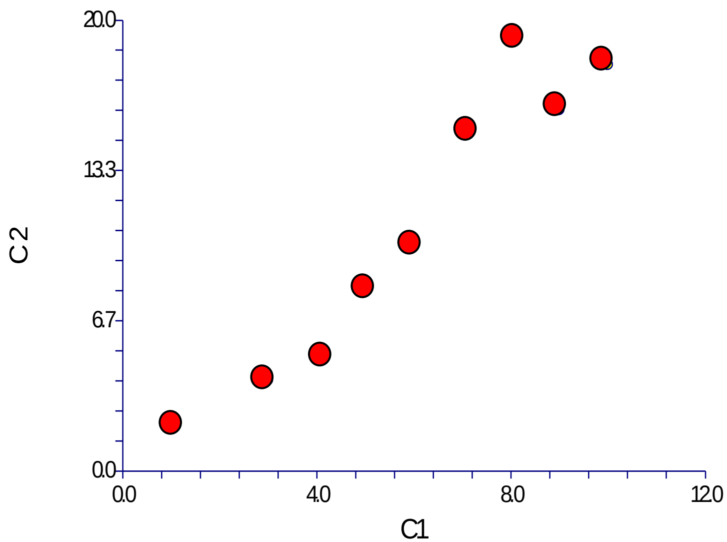
Correlation: a statistical technique that measures and describes the degree of linear relationship between two variables

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

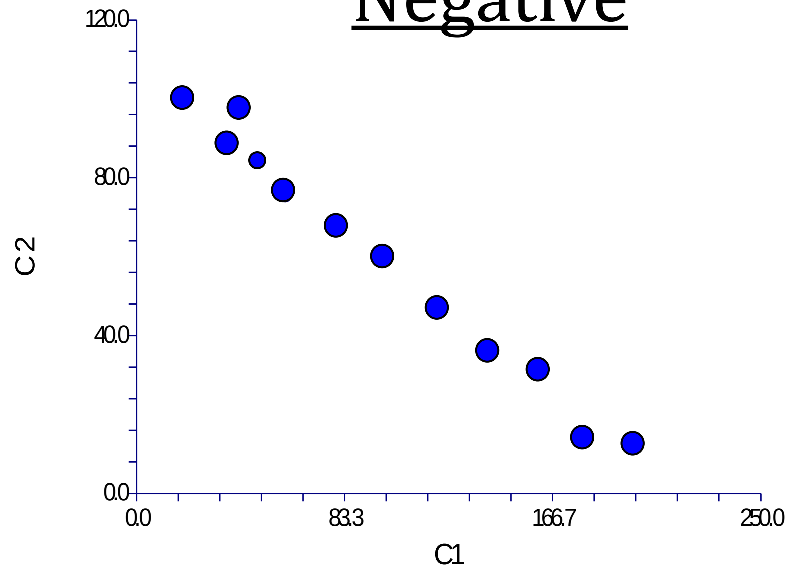
Correlation is applied on two variables

CORRELATION: 1. DIRECTION

Positive



Negative



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

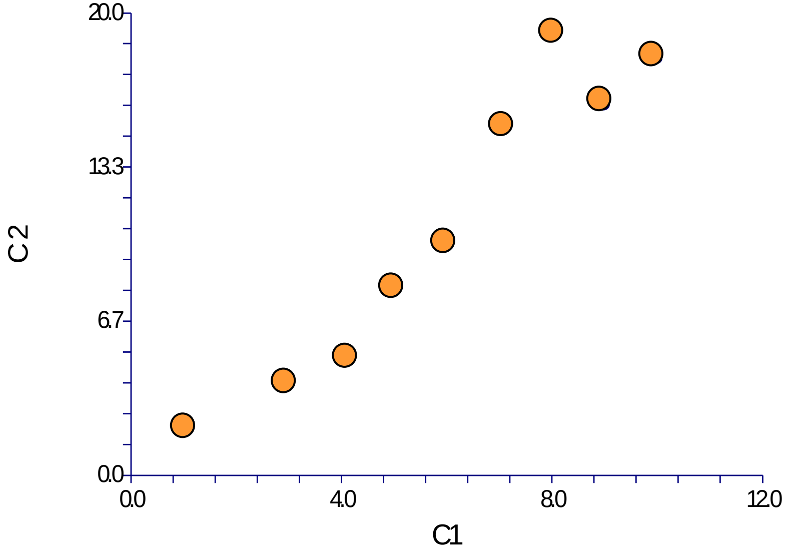
Large values of X = small values of Y
Small values of X = large values of Y

e.g. IQ (Intelligence Quotient) and SAT

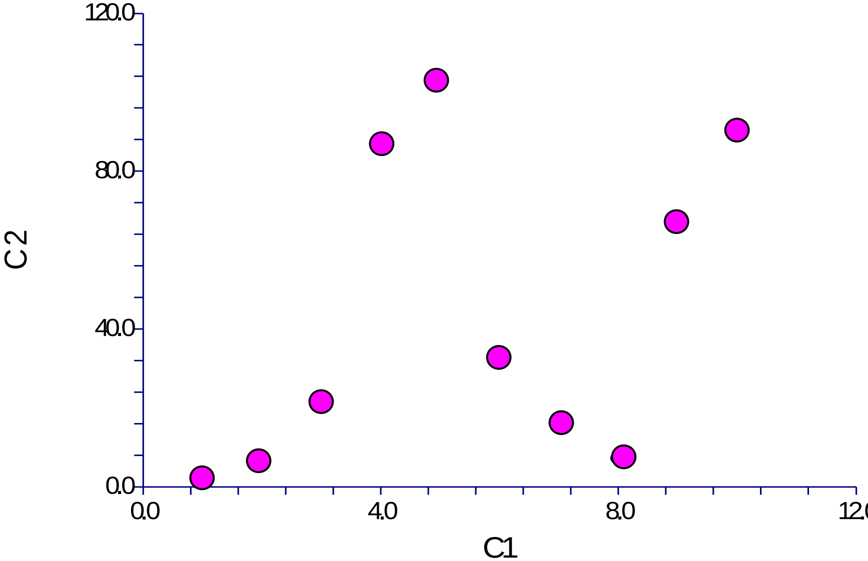
e.g. SPEED and ACCURACY

CORRELATION: 2. DEGREE OF ASSOCIATION

Strong (tight cloud)



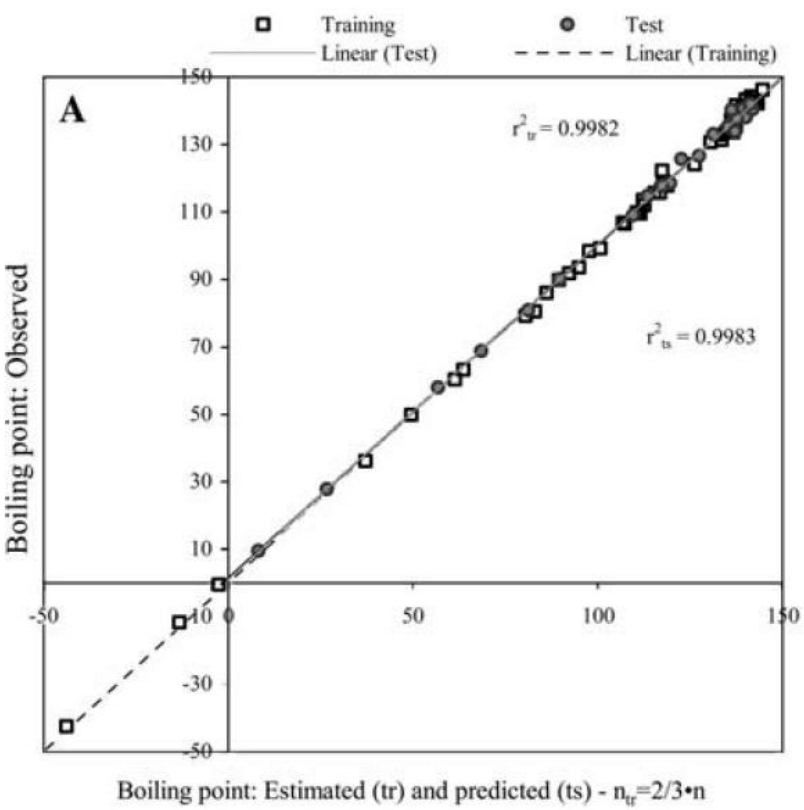
Weak (diffuse cloud)



CORRELATION: 3. FORM

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

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

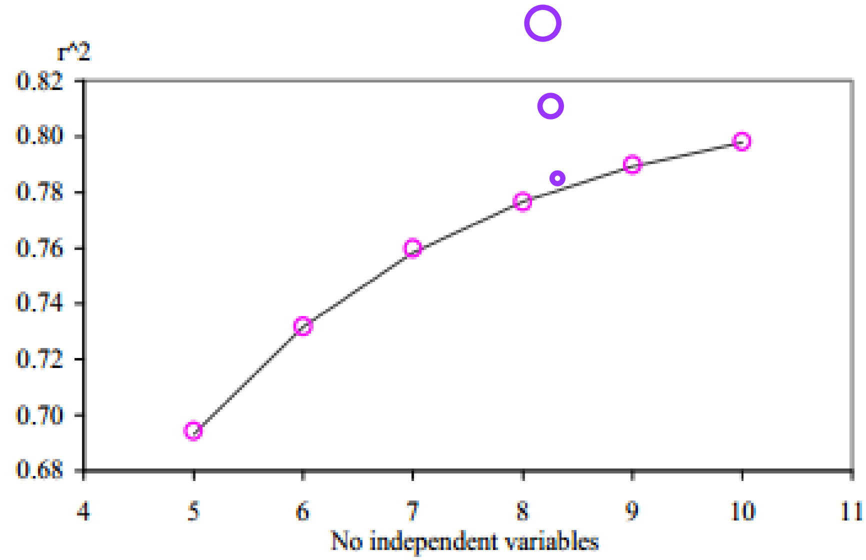


Figure 2. The dependence between r^2 and the number of independent variables for $4 < x \leq 10$

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 strength (look to the number of correlation coefficient) and direction (look to the sign of the correlation coefficient) of the linear relationship.

- Absolute value indicates strength
- +/- indicates direction

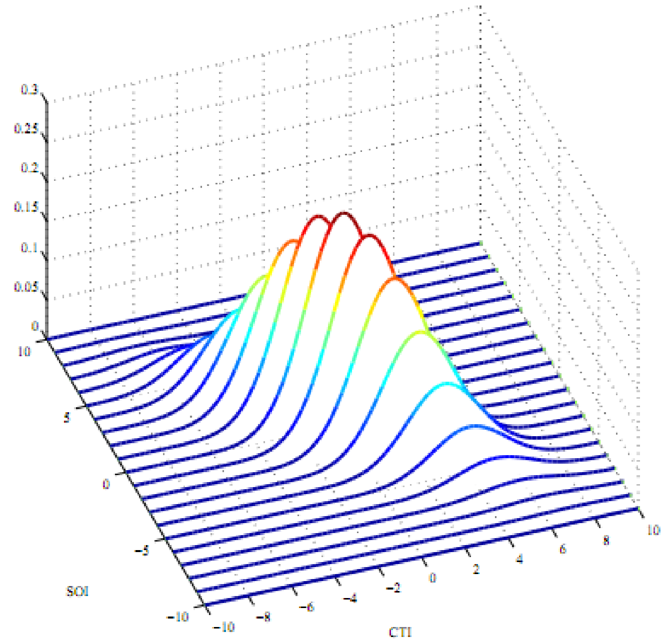
Sum of products

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

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

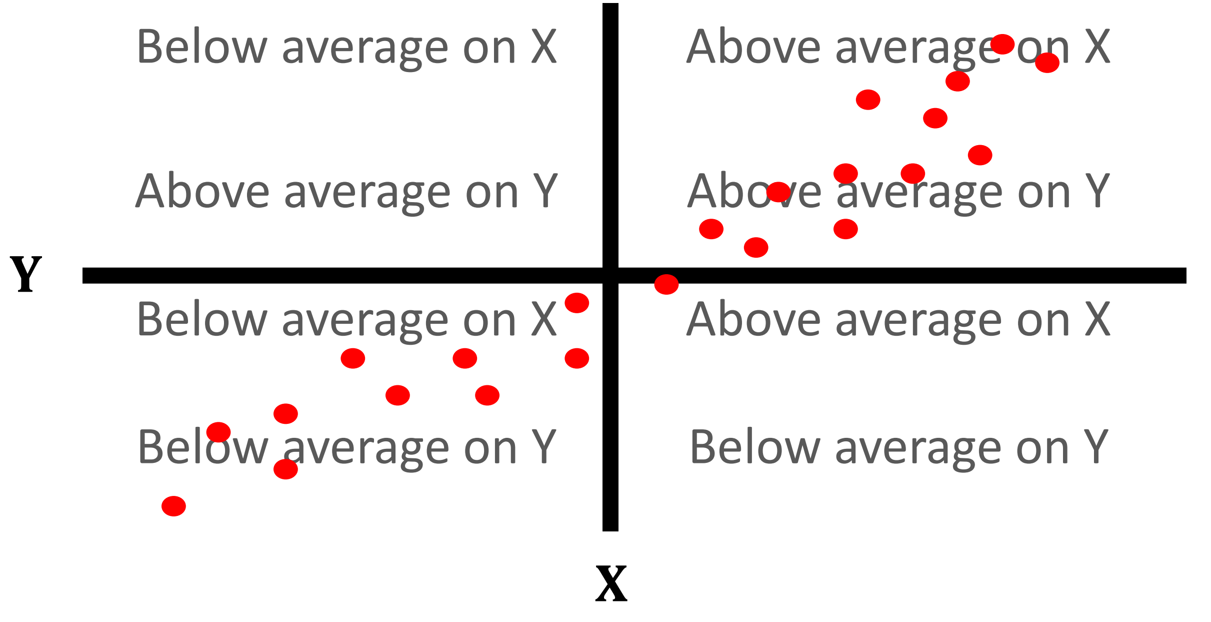


http://www.aos.wisc.edu/~dvimont/aos575/Handouts/bivariate_notes.pdf

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

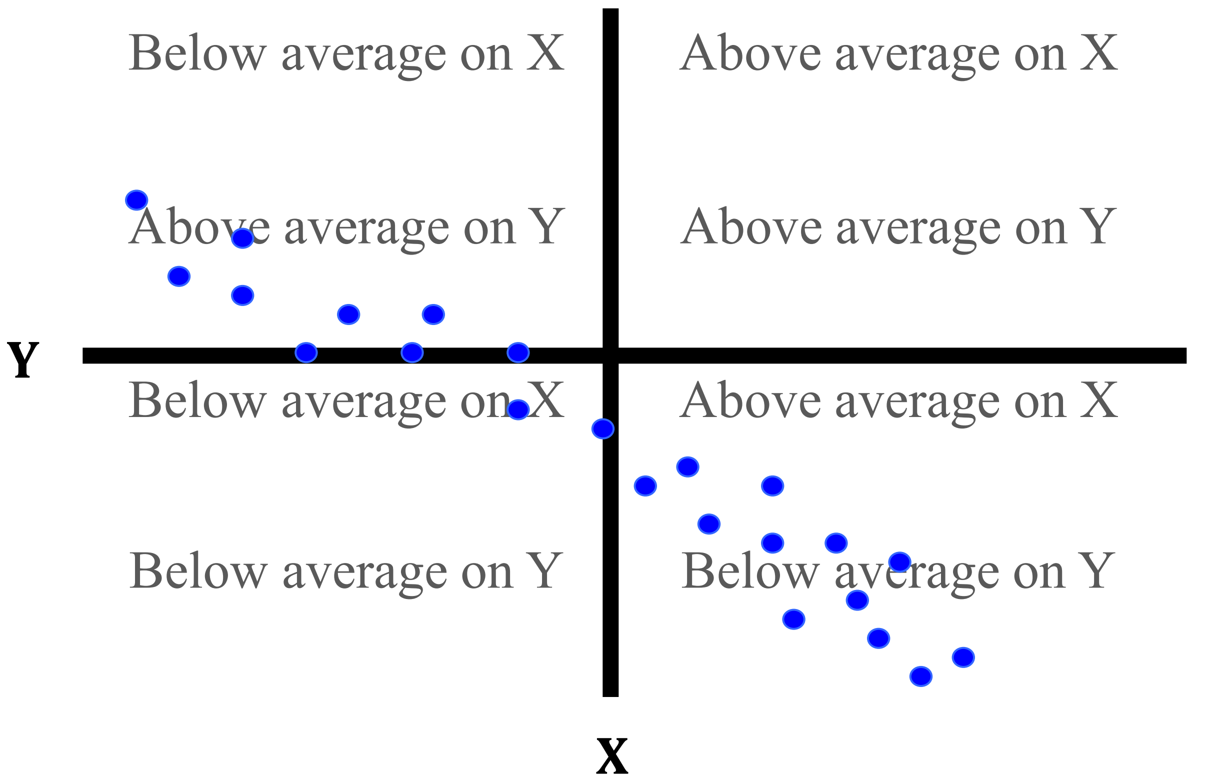
Pearson Correlation Coefficient

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



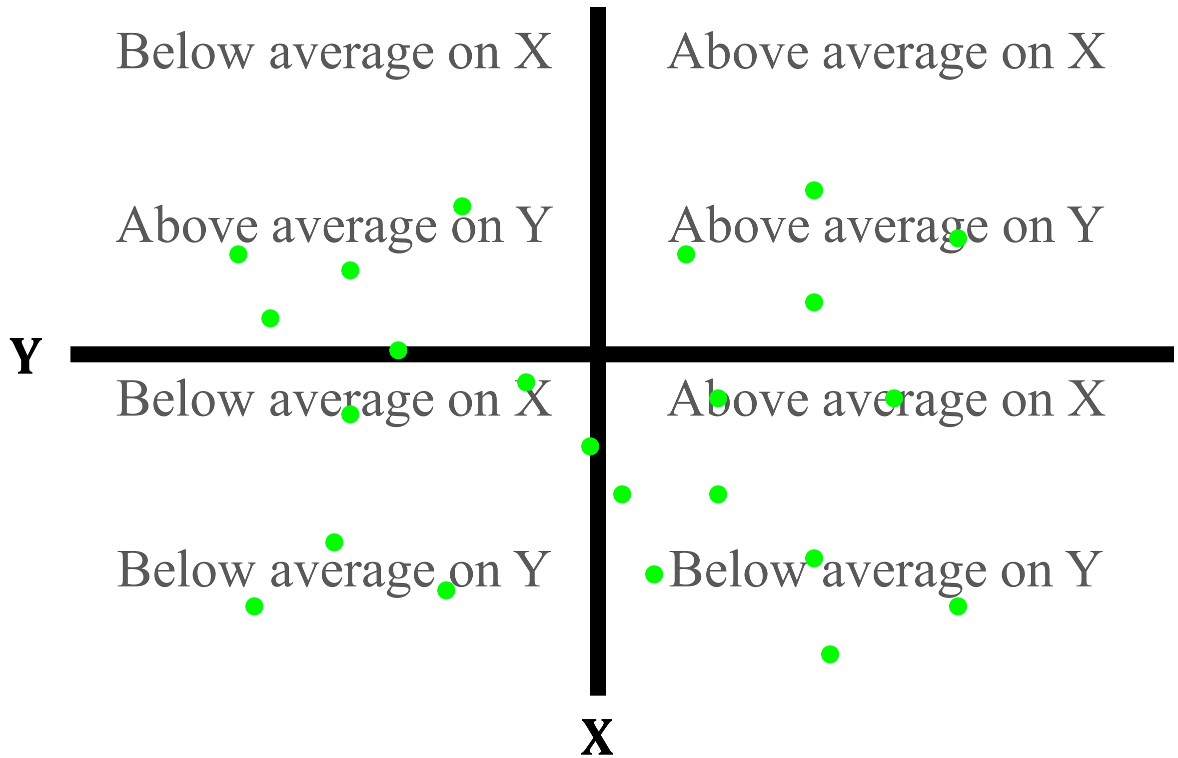
PEARSON CORRELATION COEFFICIENT

- For a strong <u>negative</u> 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?
- Empirical rules - Colton [Colton T. Statistics in Medicine. Little Brown and Company, New York, NY 1974]:
  - R ⊂ [-0.25 to +0.25] → No relation
  - R ⊂ (0.25 to +0.50] ∪ (-0.25 to -0.50] → weak relation
  - R ⊂ (0.50 to +0.75] ∪ (-0.50 to -0.75] → moderate relation
  - R ⊂ (0.75 to +1) ∪ (-0.75 to -1) → 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

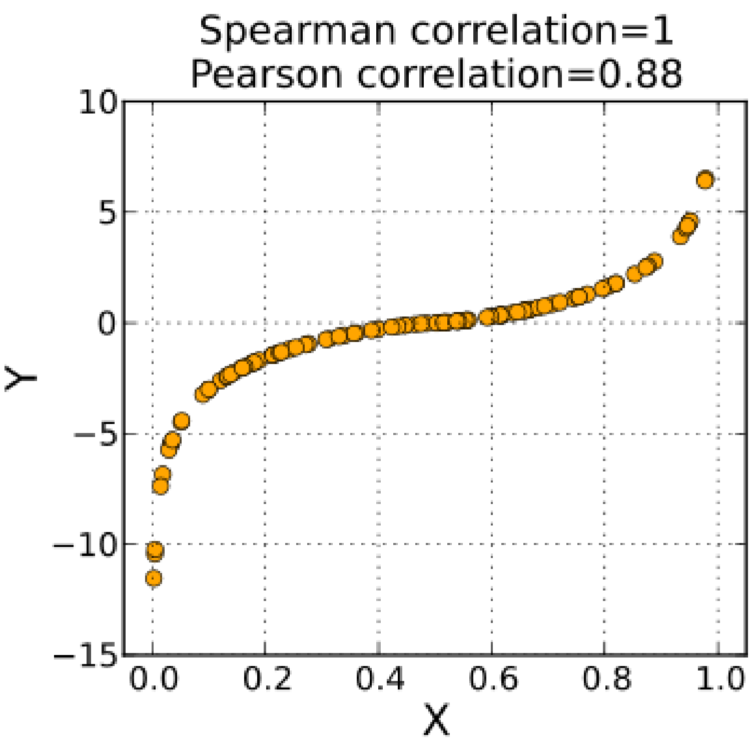


Table 3. Correlations between REACH scores and established external measures.

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

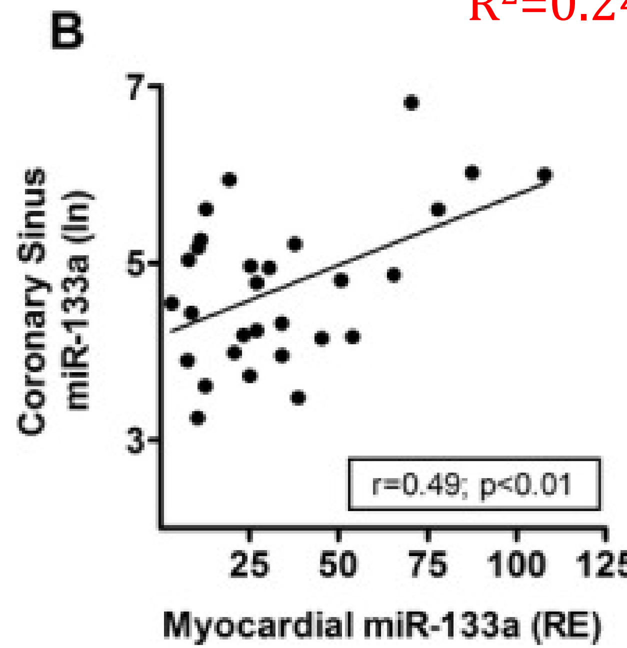
INTERPRETATION OF R-SQUARED (R^2)

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

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

- 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(no/day)	***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-stroke	Cox	ln (rate of stroke) = 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

- But how do we describe the line?
- If two variables are linearly related it is possible to develop a simple equation to predict one variable from the other
- The outcome variable is designated the Y variable, and the predictor variable is designated the X variable
- E.g. centigrade to Fahrenheit:

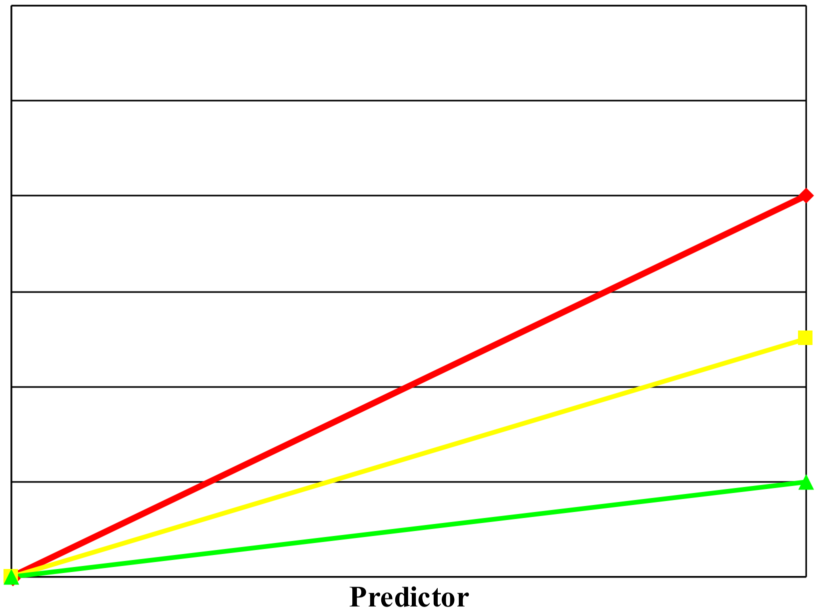
$$F = 32 + 1.8^{o}C$$

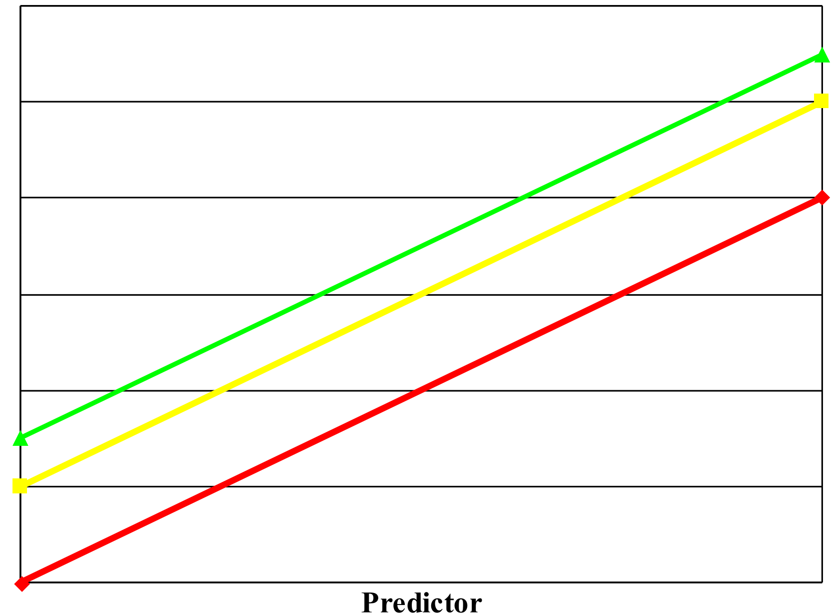
this formula gives a specific straight line

Linear Equation

- F = 32 + 1.8(C)
- General form is Y = a + bX
- The prediction equation: Y' = a+ bX
  - a = intercept, b = slope, X = the predictor, Y = the criterion
- *a and b are constants in a given line; X and Y change*

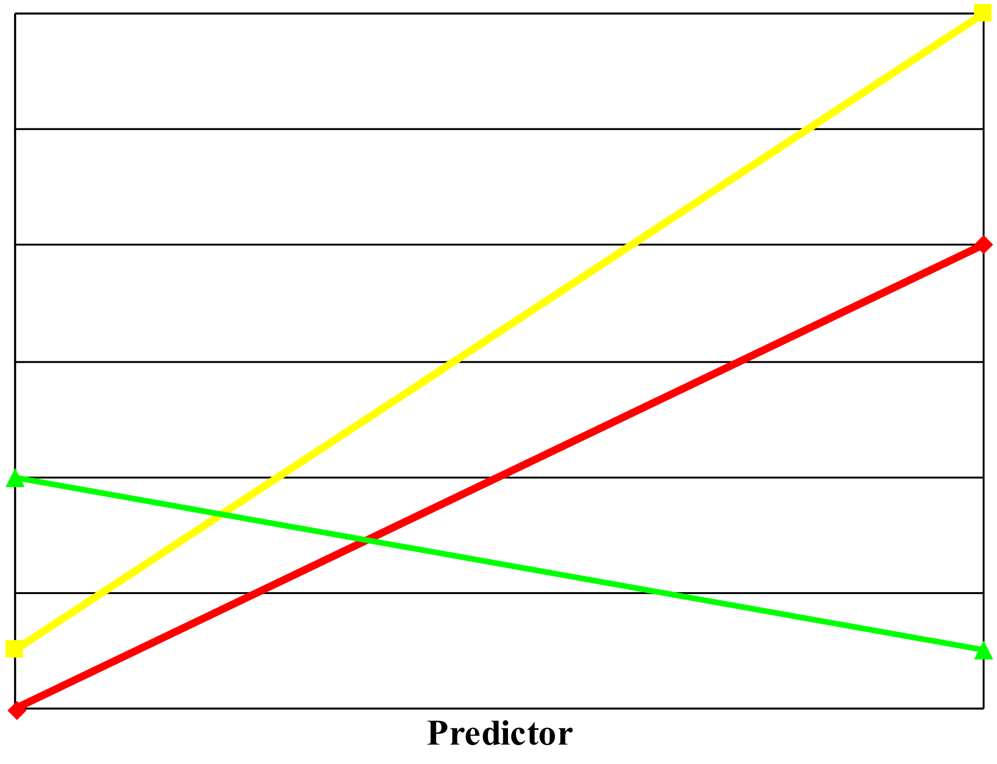
Linear Equation





Different b's… **Different a's…**

Linear Equation



Different a's and b's …

Slope and Intercept

Equation of the line: Y′= a + bX

- The slope b: the amount of change in Y with one unit change in X

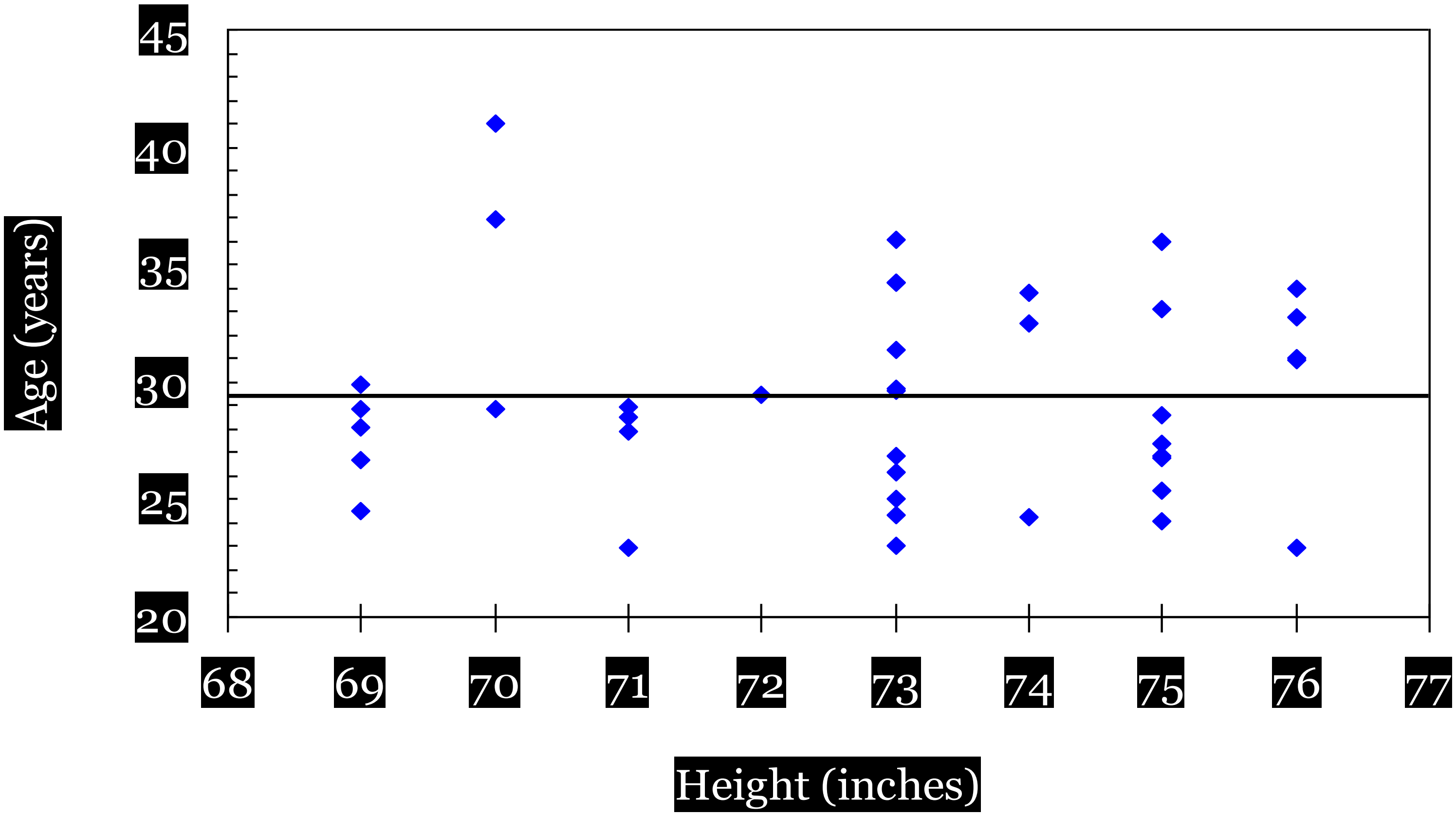
$$b = r\frac{s_y}{s_x} = \frac{SP}{SS_X}$$

- The intercept a: the value of Y when X is zero

$$a = \bar{Y} - b\bar{X}$$

- The slope is influenced by r, but is not the same as r

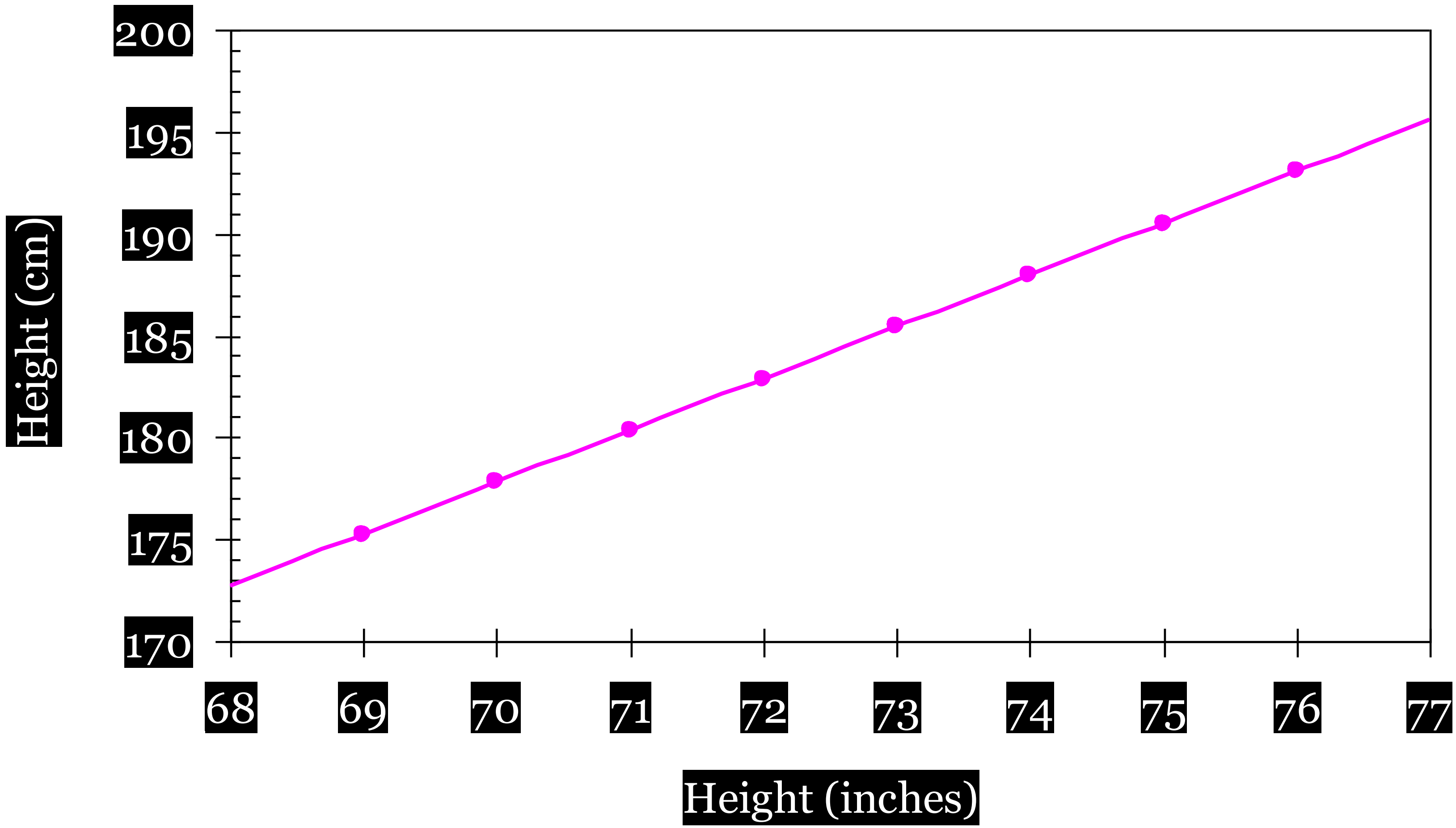
When there is no linear association (r = 0), the regression line is horizontal, b=0.



and our best estimate of age is 29.5 at all heights.

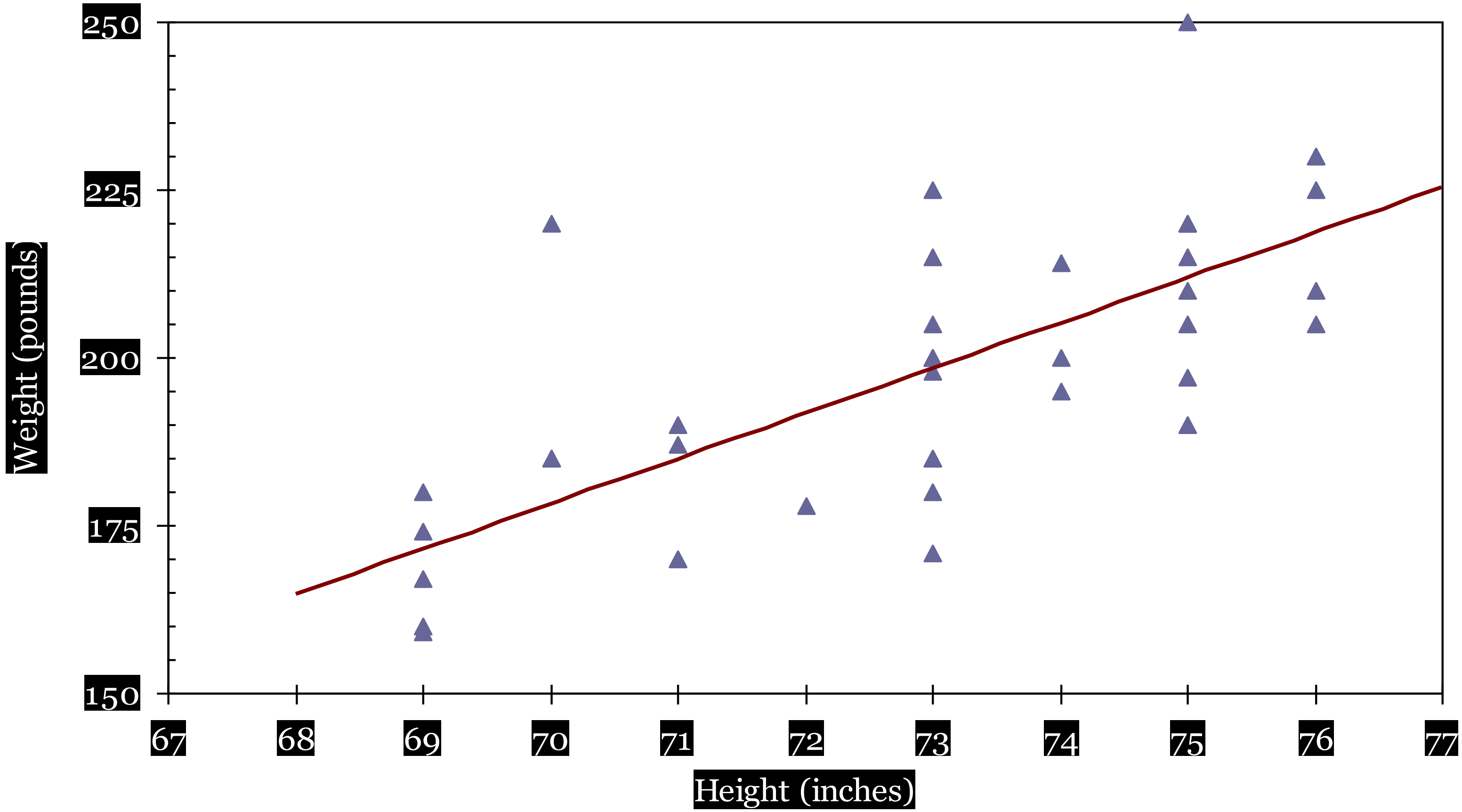
When the correlation is perfect (r = ± 1.00), all the points fall along a straight line with a slope

$$b = r\frac{s_y}{s_x}$$



When there is some linear association (0<|r|<1), the regression line fits as close to the points as possible and has a slope

$$b = r\frac{s_y}{s_x}$$

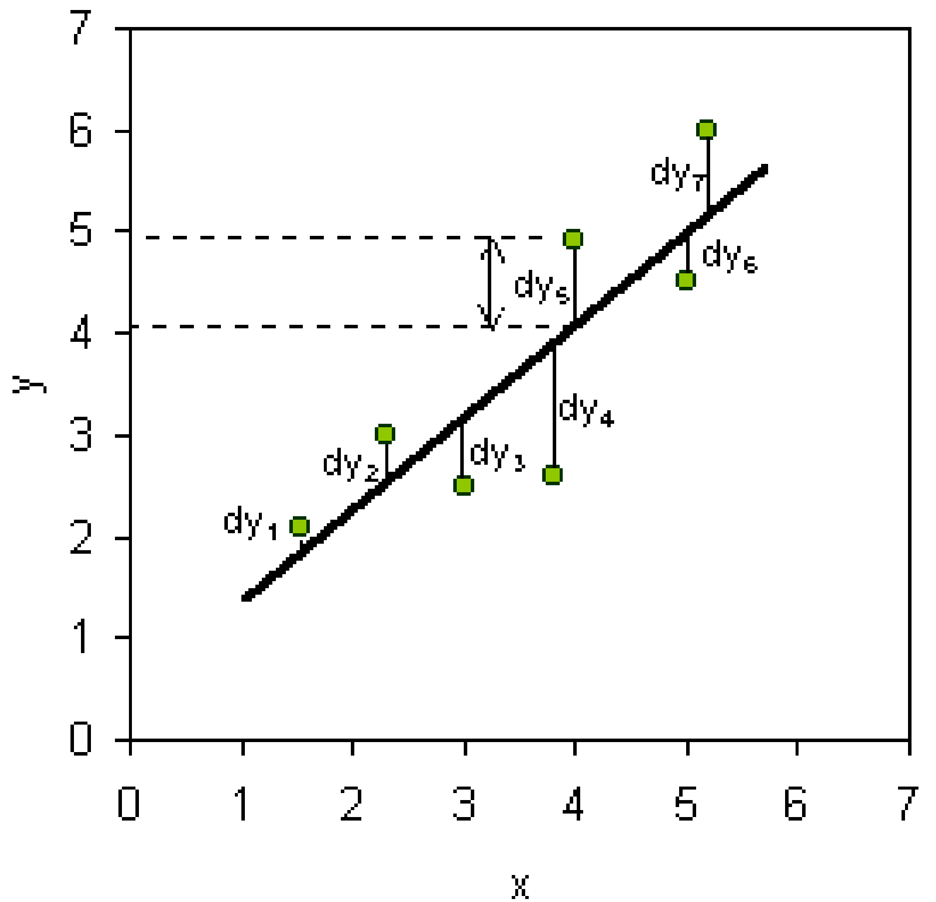


Where did this line come from?

- It is a straight line which is drawn through a scatterplot, to summarize the relationship between X and Y
- It is the line that minimizes the squared deviations $(Y' - Y)^2$
- We call these vertical deviations “residuals”

Regression Line

- Minimizing the squared vertical distances, or "residuals"



Regression Coefficients Table

Predictor	Unstandardized Coefficient	Standard error	t	p
Intercept	a	SE_a	$t=a/SE_a$	
Variable X	b	SE_b	$t=b/SE_b$	

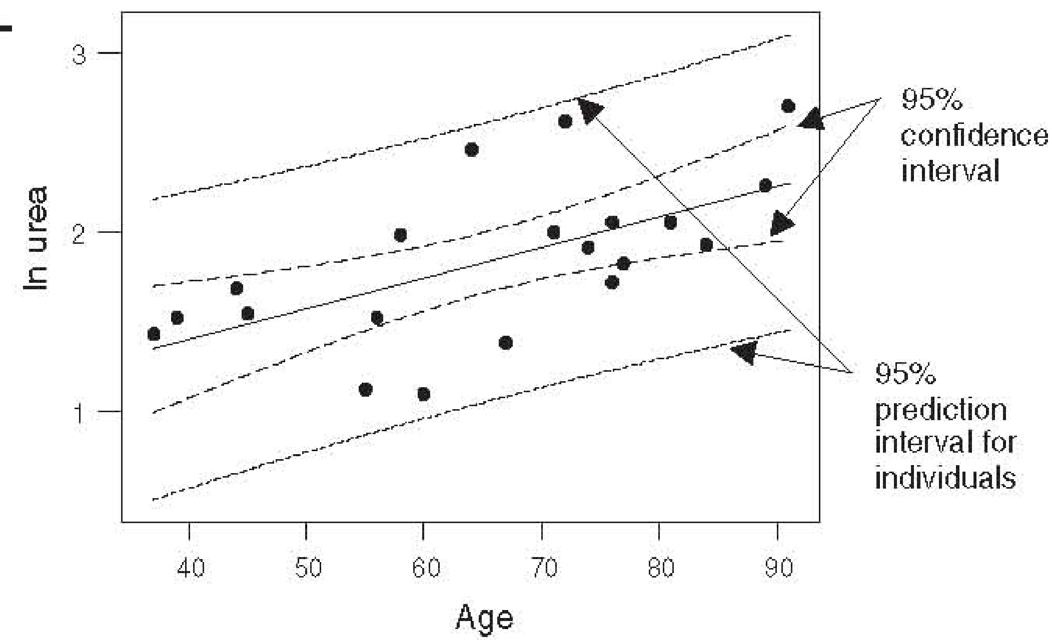
Regression parameter estimates, *P* values and confidence intervals for the accident and emergency unit data

	Coefficient	Standard error of coefficient	t	*P*	Confidence interval
Constant, or intercept	0.72	0.346	2.07	0.054	−0.01 to +1.45
ln urea	0.017	0.005	3.35	0.004	0.006 to 0.028

Linear Regression

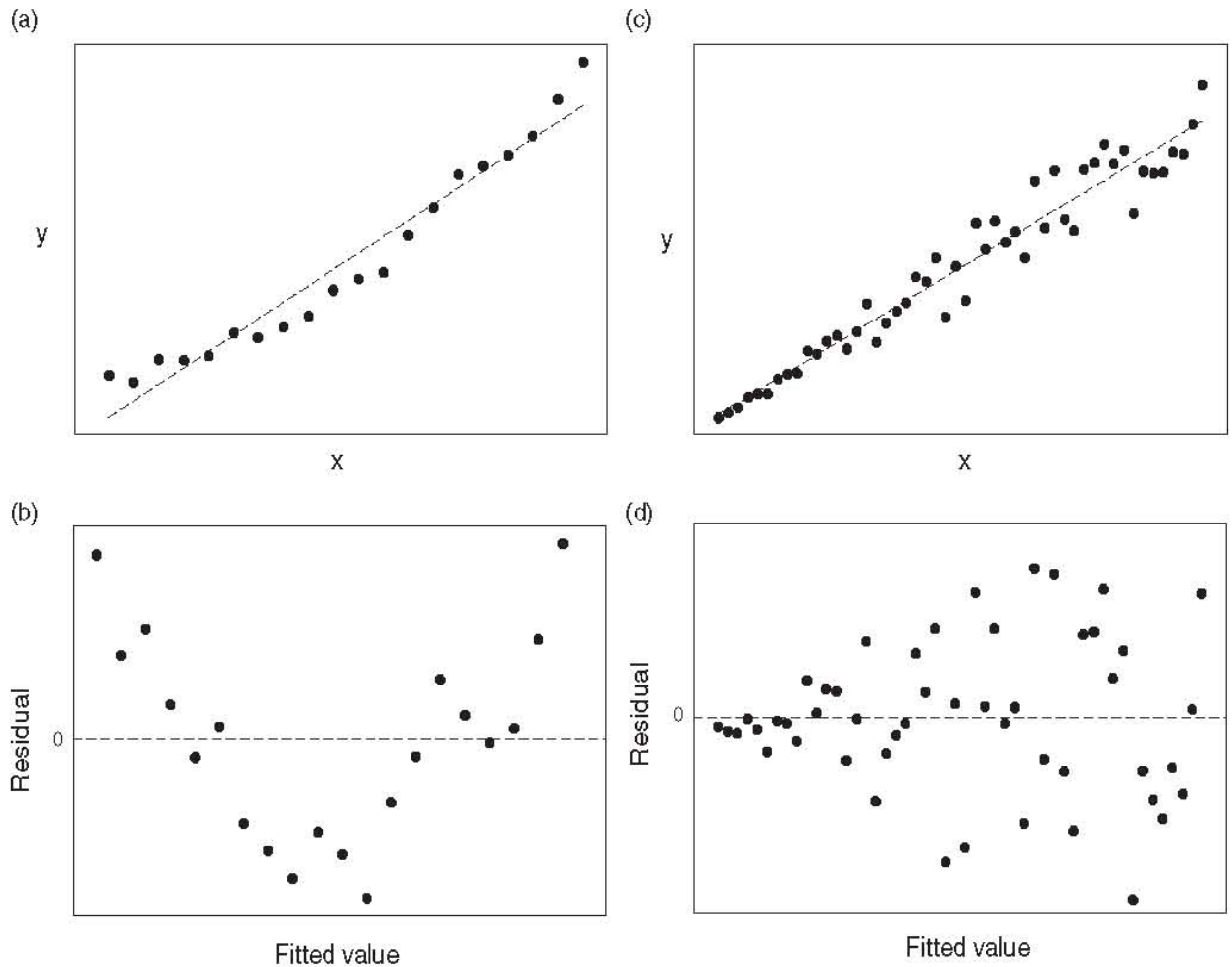
Analysis of variance for the accident and emergency unit data

Source of variation	Degrees of freedom	Sum of squares	Mean square	F	*P*
Regression	1	1.462	1.462	11.24	0.004
Residual	18	2.342	0.130		
Total	19	3.804			



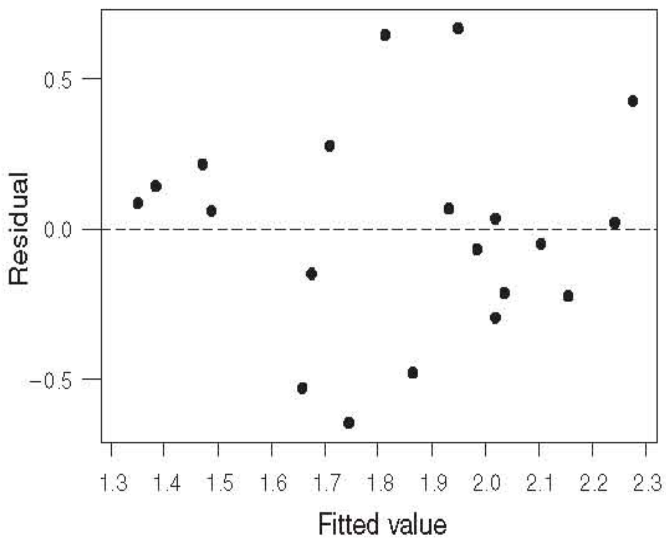
Regression line, its 95% confidence interval and the 95% prediction interval for individual patients.

Linear Regression

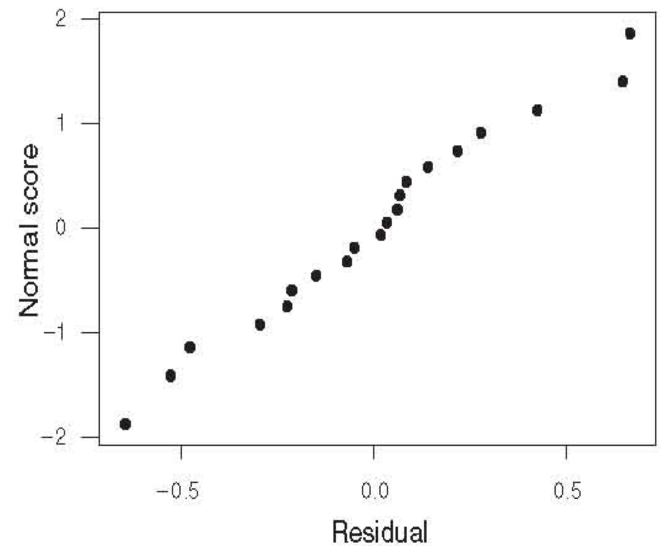


(a) Scatter diagram of y against x suggests that the relationship is nonlinear. **(b)** Plot of residuals against fitted values in panel a; the curvature of the relationship is shown more clearly. **(c)** Scatter diagram of y against x suggests that the variability in y increases with x. **(d)** Plot of residuals against fitted values for panel c; the increasing variability in y with x is shown more clearly.

Linear Regression



Plot of residuals against fitted values for the accident and emergency unit data.



Normal plot of residuals for the accident and emergency unit data.

LINEAR REGRESSION MODEL

http://www.sciencedirect.com/science/article/pii/S2213158214001648

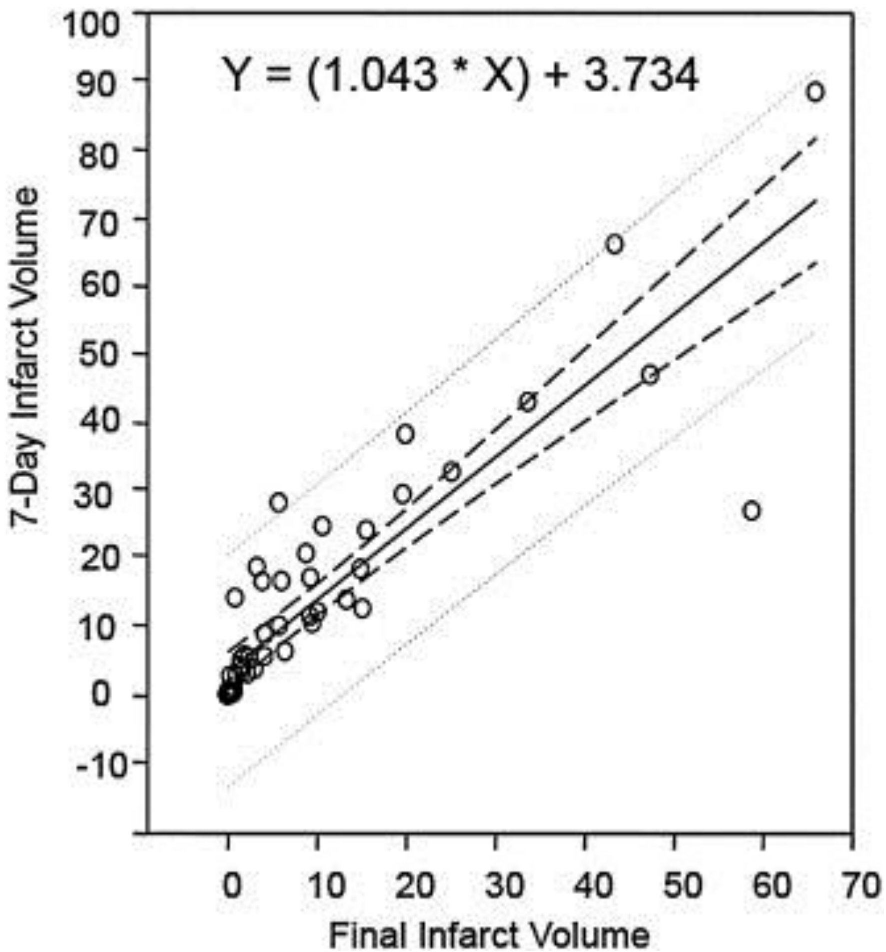


Fig. 4.
Linear regression analysis of 7-day and final infarct volumes. The solid line represents the regression line and dashed lines represent the 95% prediction and confidence intervals. Calculated *p* values for slope (1.043) and intercept (3.734) were <0.001 and 0.009, respectively.

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

Thank you.

THINGS GOT REALLY INTERESTING WHEN THE STATISTICIAN STARTED DOING WARD ROUNDS

"Our statistician will drop in and explain why you have nothing to worry about."