Simple Linear Regression

Instructor

Michał Kowalewski
Florida Museum of Natural History
University of Florida
kowalewski@ufl.edu

Correlation versus Regression

Correlation

Aims to assess if inter-dependence exists between two variables. Correlation analysis does not aim to explore causal relations or develop predictive models.

Typically, r (or occasionally r^2) and associated statistical measures (t, p) are the primary numerical estimates reported in the correlation analysis.

In simple bivariate case, both variables are treated equally and order of variables does not matter:

cor(x,y) = cor(y,x)

Regression

A modelling effort that attempts to find the best model for linking a response variable to one (or multiple) explanatory variables. These models can also be used to develop predictors.

Regression analysis focuses on model parameters and r^2 is used (for some types of regression modeling) to assess proportion of variance in response variable accounted for by explanatory variable.

Variables are typically partitioned into one response and one or more explanatory variables. The order of variables does matter:

lm(x ~ y) ≠ lm(y ~ x)

Correlation and Regression

(*r* versus r^2)

The *coefficient of determination* $\boldsymbol{r^2}$ is equivalent to square of the parametric Pearson's product-moment correlation coefficient ***r*** correlation. For this and other reasons, regression and correlation are often confused with each other.

Pearson's Correlation *r*

- Pearson's product-moment correlation coefficient
- Varies from -1 to 1
- ***r*** is routinely reported in correlation analysis
- Parametric test for $r = 0$ is based on t statistics (as discussed previously)

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})}}$$

Coefficient of Determination r^2

- $\boldsymbol{r^2} = S^2_{\hat{Y}} / S^2_Y$ $S^2_{\hat{Y}}$ – variance of predicted (or fitted) values; S^2_Y – variance of Y
- $\boldsymbol{r^2}$ varies from 0 to 1
- $\boldsymbol{r^2}$ is routinely reported in regression analysis
- $\boldsymbol{r^2}$ computed as ratio of two variances is equal to the square of Pearson's product-moment correlation coefficient ***r***
- Parametric test for $\boldsymbol{r^2} = 0$ is based on ***F*** statistics (ratio of variances)
- Measures amount of variability in one variable accounted by correlating that variable with the second variable

Correlation and Regression

(examples)

Correlation

Example 1: Pearson

- Are Phanerozoic global biodiversity and rock outcrop area inter-related? H[0] ***r*** = 0 (detrended data)
- How strong is this interrelation? ***r***

Example 2: Spearman

- Are rank abundances of species in live and death assemblage similar? H[0] ***rho*** = 0
- How similar? ***rho***

Regression

Example 1 (Type I regression problem):

- What is the effect of ambient storage temperature on amino-acid ratios in fossil skeletons stored in a museum (a heating lab experiment in various temperature settings)?
- How strong is that effect? r^2
- Can we estimate ("predict") the original ratios?

Example 2 (Type II regression problem):

- Are skull length and limb length allometrically related?
- How strong is that inter-relation? r^2
- What is the nature of this inter-relation: negative allometry, positive allometry, isometry?

Basic Terminology of Regression

In simplest bivariate case, regression deals with two variables. For Type I regression they can be explicitly sorted into:

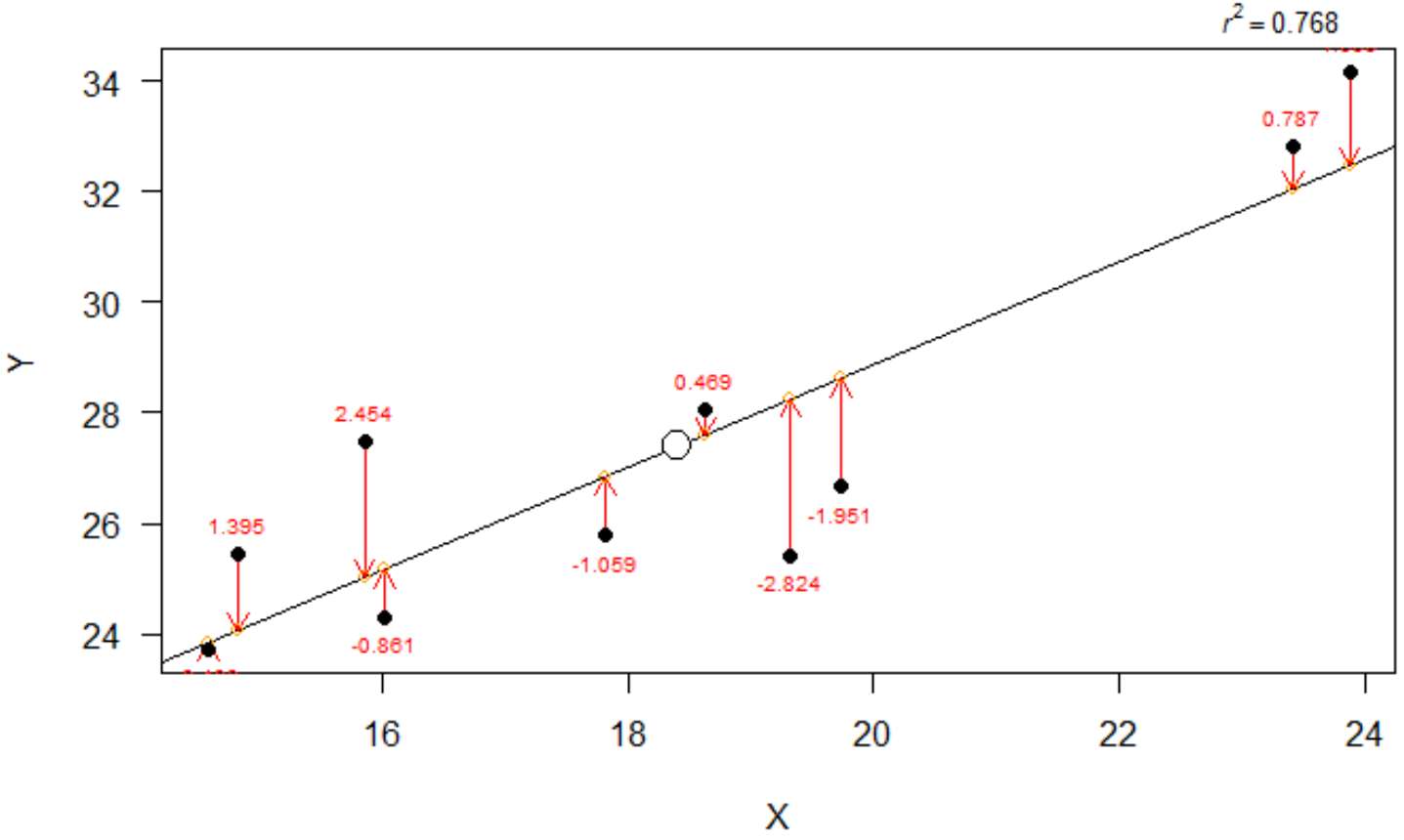
- The independent (regressor or predictor) variable.
- The dependent (criterion) variable.

Linear regression - Assumes linear dependence between the variables. It is directly related to Pearson's correlation (r^2 = square of Pearson's correlation r, but measures the amount of variability in the dependent variable accounted for by regression on independent variable)

Simple regression - Denotes bivariate analysis whereas multiple regression denotes analysis when a single dependent variable is evaluated in terms of 2 or more independent variables.

Simple linear regression - The most popular (and simplest) analysis which denotes bivariate analysis measuring dependence as a linear function y = alpha + betaX + error.

Model 1: Ordinary Least Squares (OLS)



$\hat{Y} = a + b*X$ *OLS regression model is defined by two parameters (a – intercept, b – slope)*

The OLS line minimizes residuals in the direction of Y

Individual residuals, which can be negative or positive, are computed as follows:

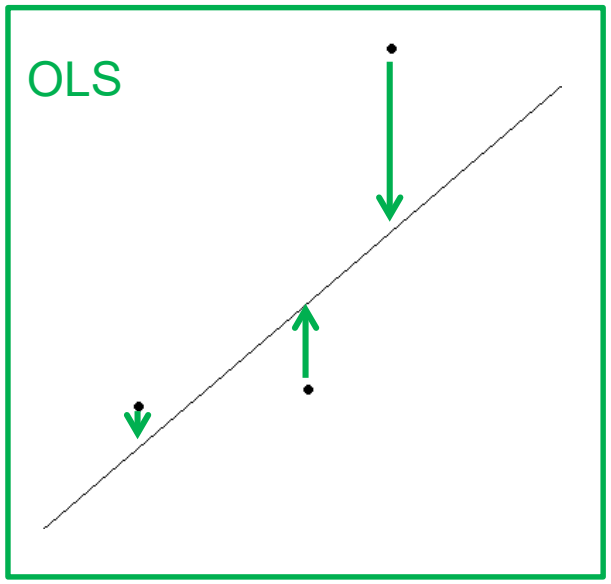
${}^{R}Y_i = \hat{Y}_i - Y_i$ $\hat{Y}_i$ – predicted value of ith observation, Y_i – observed value of ith observation

Regression: Model I versus Model II

Model I

Y is dependent and X is independent
X is known/controlled (no error)
(or X much smaller errors than Y)

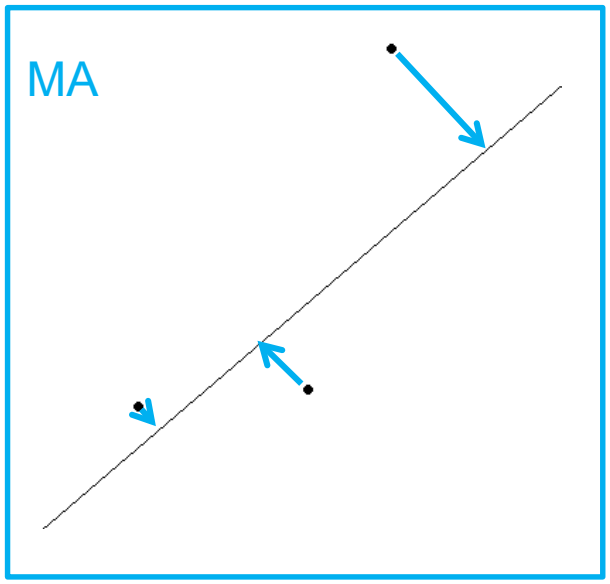
Ordinary Least Squares



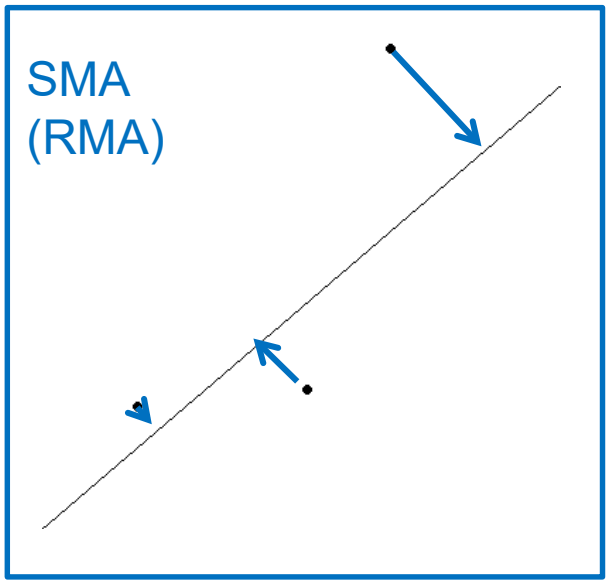
Model II

X and Y are inter-dependent
and/or X and Y have comparable errors

Major Axis Regression (MA)



Standardized (or Reduced) MA

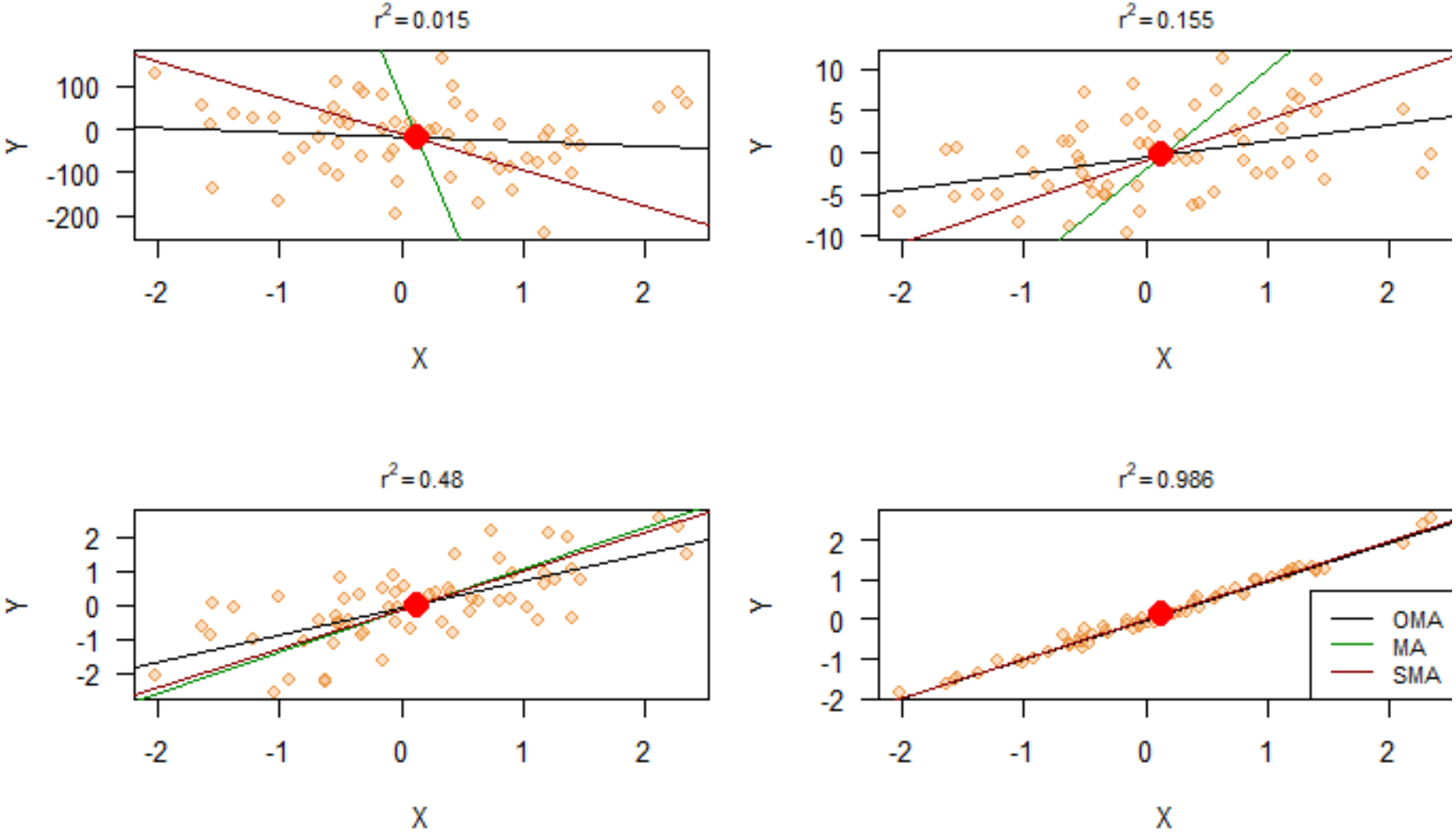


Regression: Model I versus Model II

Models	slope parameter *b*	intercept *a*
Model I OLS	$b_{OLS} = \frac{S_{xy}}{S_x^2}$	a_{OLS} = mean(*Y*) – b_{OLS}*mean(*X*)
Model II MA	$b_{MA} = \frac{S_y^2 - S_y^2 + \sqrt{(S_y^2 - S_y^2)^2 + 4(Sxy)2}}{2S_{xy}}$	a_{MA} = mean(*Y*) – b_{MA}*mean(*X*)
Model III SMA	$b_{SMA} = \frac{sign(S_{xy}) * S_y}{S_x}$	a_{SMA} = mean(*Y*) – b_{SMA}*mean(*X*)

r and ***r*** **²** are the same for all models

NOTE: Package {lmodel2} in *R* produces (among others) OLS, MA, and SMA



- OLS and MA models converge when correlations are strong
- Models diverge as r2 approaches 0
- However, all models pass through the centroid

Significance Testing in Simple Linear Regression

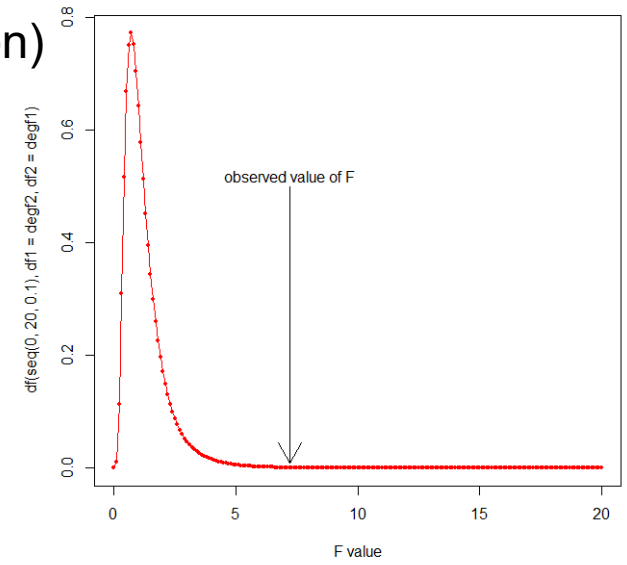
r^2

Because r^2 can be thought as ratios of variances (total versus residual), a ratio-variance density distribution F is used in parametric tests.

Significance of R^2 can be estimated using various formulas.
A simple way to compute F is as follows:

$$F = \frac{r^2}{1 - r^2} \frac{DF_2}{DF_1}$$

with degree of freedoms determined by number of predictors *k*:
$DF_1 = k = 1$ (for simple linear regression)
$DF_2 = n - k - 1$ (n – 2 for simple linear regression)



Significance Testing in Simple Linear Regression

b – slope

We may want to know if b is different from 0: H[0]: $\beta = 0$
We may also want to know if b is different from 1: H[0]: $\beta = 1$
Or some other slope value: H[0]: $\beta = X$

The parametric t-test is based on the following formula:

$$t = \frac{b - b_0}{S_b}$$

b – observed slope estimate
b_0 – slope postulated by null hypothesis
S_b – standard error of slope estimate