Co-financed by the Connecting Europe
Facility of the European Union

xAIM

Linear Regression Models

Trustworthy AI - Lecturer: Emanuela Raffinetti; Python instructor: Alex Gramegna
E-mail: emanuela.raffinetti@unipv.it; alex.gramegna01@universitadipavia.it

Co-financed by the Connecting Europe Facility of the European Union

1. Correlations
   - Simple Correlation
   - Partial Correlation

2. Regression
   - Simple linear regression
   - Multiple Linear Regression

Recap on the correlation coefficient - I

- The correlation coefficient is a measure of the direction and strength of the linear relationship between two quantitative variables.
- It is calculated using the covariance between X and Y and the standard deviation of both variables:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{(\sigma^2(X))}\sqrt{(\sigma^2(Y))}}$$

- Note: The correlation coefficient is a measure of linear relationship! Two variables can exhibit a non-linear relationship.

Recap on the correlation coefficient - II

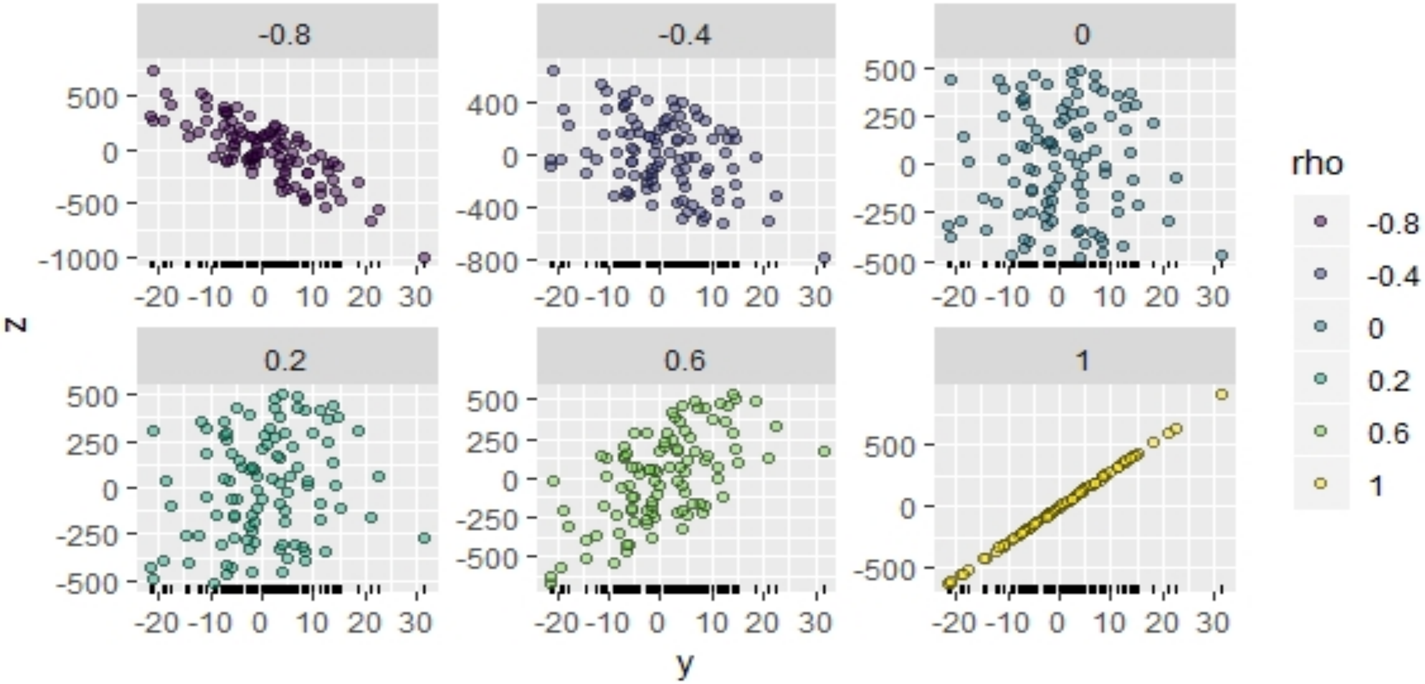


Figure: Varying correlation coefficient

Recap on the correlation coefficient - III

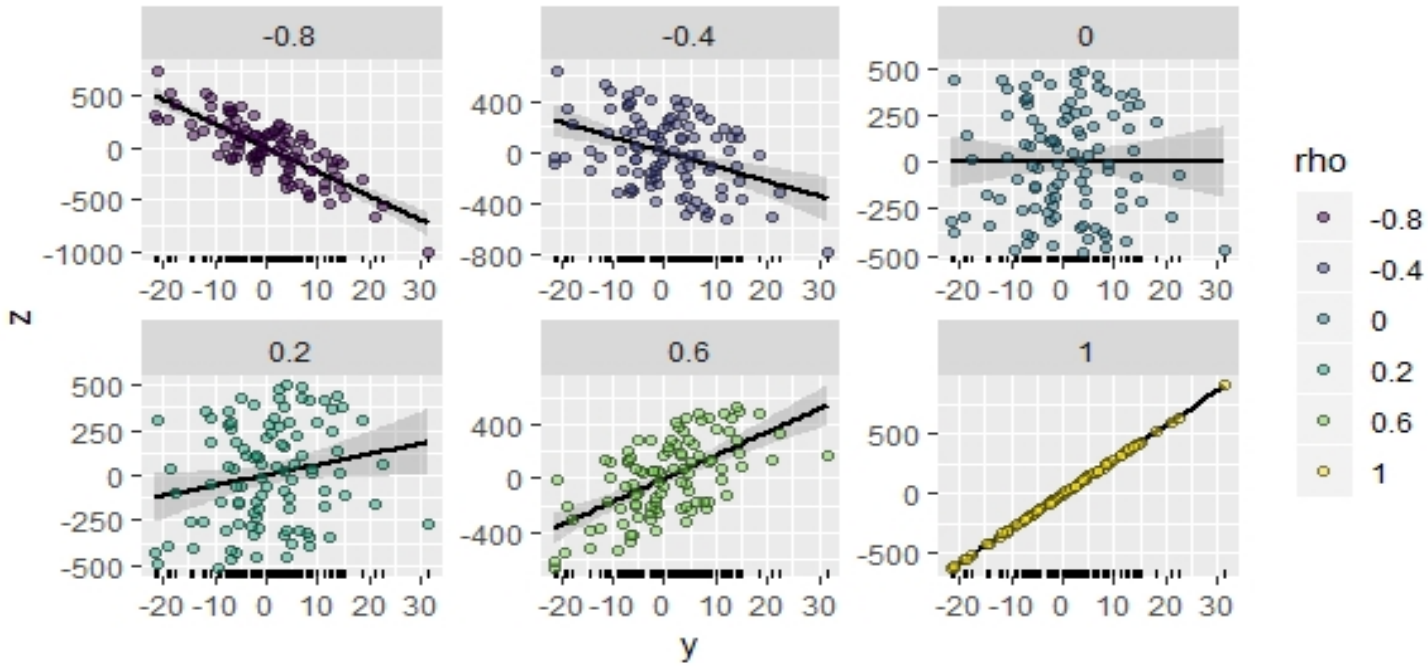


Figure: Varying correlation coefficient [smoother = linear model]

Recap on the partial correlation coefficient - I

- Partial correlation is a measure of the strength and direction of a linear relationship between two continuous variables whilst controlling for the effect of one or more other continuous variables.
- Partial correlation measures the correlation between X and Y, controlling for Z.
- Comparing bivariate correlation to partial correlation allows to determine if the relationship between X and Y is direct or indirect (spurious).

Recap on the partial correlation coefficient - II

- The partial correlations can be directly obtained from the inverse of a variance-covariance matrix.
- Let X represent a set of response variables and Σ denote a variance-covariance matrix:

$$X \sim N(O, \Sigma)$$

$$\Theta = \sum^{-1}$$

- The partial correlation between X_i and X_j conditioned upon all other X will be:

$$Cor(X_i, X_j | X_{-(i,j)}) = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

Introduction

- Linear regression is one of the most popular and widely used statistical modeling approach.
    - Transparent and relatively easy to understand
    - Highly robust to statistical anomalies
    - It provides a basis for more complex methods (such as neural networks)

Specification – Simple Linear regression

- The simple linear regression model

$$y_i = \alpha + \beta x_i + e_i$$

  - y_i = dependent variable (outcome variable, response variable, explained variable, predicted variable)
  - x_i = independent variable (explanatory variable, control variable, predictor variable, regressor, covariate)
  - e_i = error term (noise, disturbance)
  - α = intercept parameter
  - β = slope parameter

- Estimated Regression Line:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

Graphical representation

$$\varepsilon_i = y_i - \hat{y}_i$$

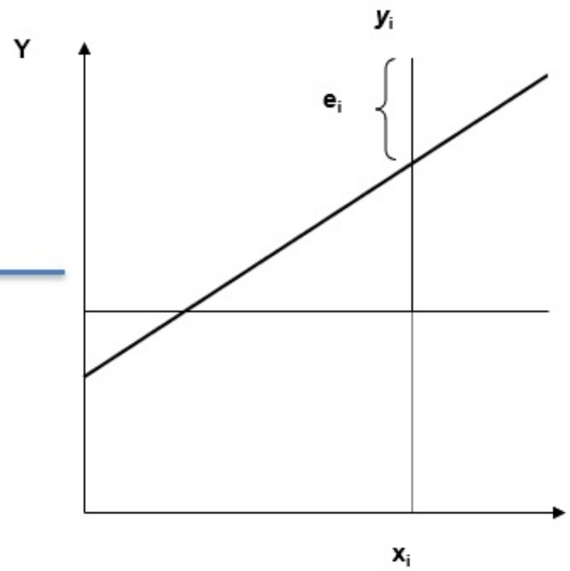


Figure: Simple Linear Regression

Calculation of coefficients

- How to estimate the slope:

$$\hat{\beta} = \frac{\sum(x_i - \bar{y})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

- How to estimate the intercept:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

- Interpretation: $\hat{\beta}$ gives us the change in Y if X changes by one unit; $\hat{\alpha}$ gives us the predicted value of Y if X is equal to zero.

Goodness of fit - I

- R-squared defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

where

SSR = Regression Sum of Squares = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y}\ )^2$

SSE = Sum of Squared Errors = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

SST = Total Sum of Squares = $\sum_{i=1}^{n}(y_i - \bar{y}\ )^2$

Co-financed by the Connecting Europe Facility of the European Union

Goodness of fit - II

- The goodness of fit can also be investigated through the F-test.

$$f = \frac{R^2}{(1-R^2)/(n-2)}$$

  where f follows an F distribution with 1 and $n-2$ degrees of freedom.

- The practical interpretation is that a bigger R^2 will lead to high values of the F statistic, so if R^2 is big (which means that a linear model fits the data well), then the corresponding F statistic should be large.

Co-financed by the Connecting Europe Facility of the European Union

T-test

- In simple linear regression, performing an F-test is equivalent to testing the hypothesis that the slope (β) is equal to 0:

$$f = t^2$$

  where t follows a Student's T distribution with $n-2$ degrees of freedom.
- With the F-test or the T-test we verify whether changes in the explanatory variable are significantly associated with changes in the response.
- Once the p-value is calculated, the null hypothesis is rejected for any $\alpha > p-value$, while the null hypothesis is not rejected when $\alpha < p-value$.

Co-financed by the Connecting Europe Facility of the European Union

Assumptions

- Linearity - the relationship between Y and X is linear
- Normality - residuals are normally distributed - necessary for inferential results
- Homoskedasticity - the variance of errors are similar across the values of the independent variable
- No auto-correlation - residuals are independent from each other

Homoskedastic

- Uniform spread of errors around the regression line.

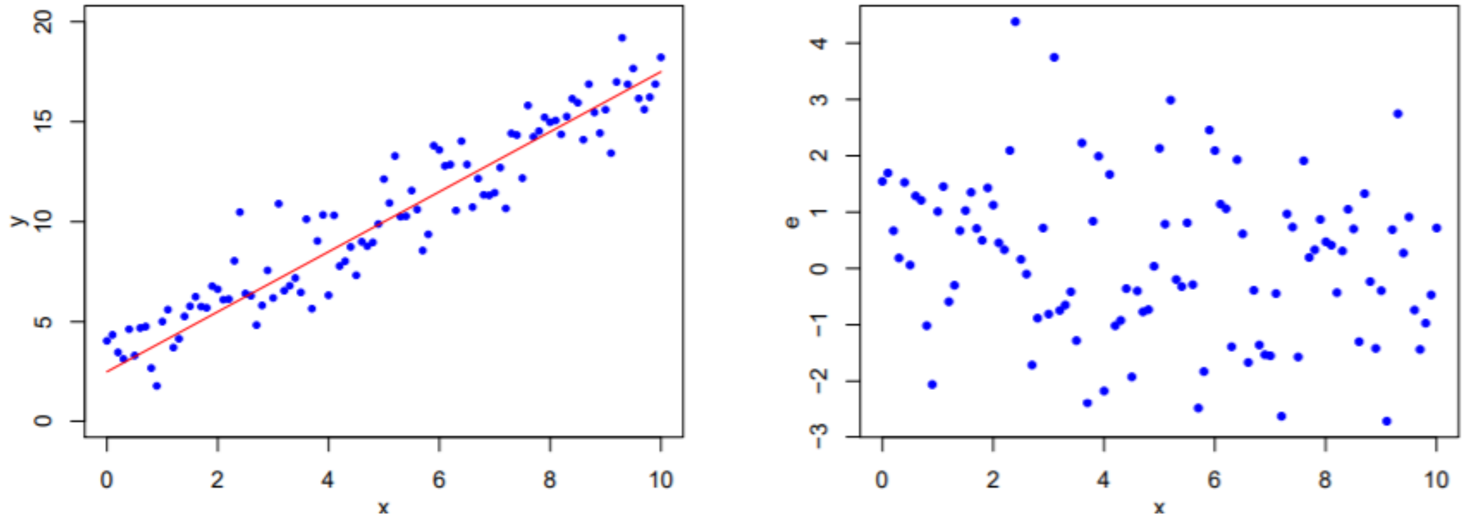
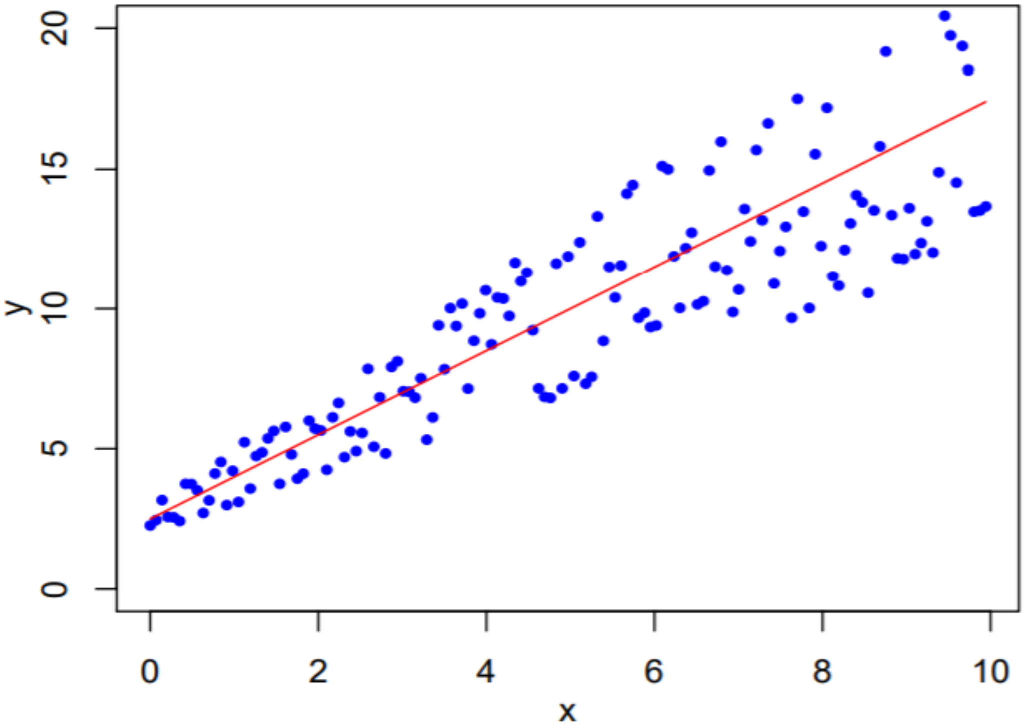


Figure: Left: regression line; Right: residuals

Heteroskedastic

- Errors are not uniformly spread around the regression line.



From Simple to Multiple Linear Regression

- In a simple linear regression model, a single response measurement Y is related to a single predictor (covariate, regressor) X for each observation.
- In most applications, more than one predictor variable is available. This leads to the following "multiple regression" function:

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \ldots + \beta_p x_{i,p} + e_i$$

- Goodness of fit can be derived as before.
- Similar key assumptions, plus weak collinearity (weak dependence between the explanatory variables).

Calculation of coefficients

- In order to estimate β, we use a least squares approach that is analogous to what we did in the simple linear regression case. That is, we want to minimize

$$\sum_i (y_i - \alpha - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2$$

  over all possible values of the intercept and slopes.

- This is obtained by setting

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Goodness of fit - I

- As in the simple linear model, we have the SST = SSE + SSR decomposition:

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

- The sums of squares with degrees of freedom:

Source	Formula	DF
SSTO	$\sum(Y_i - \bar{Y})^2$	$n-1$
SSE	$\sum(Y_i - \hat{Y}_i)^2$	$n-p-1$
SSR	$\sum(\hat{Y}_i - \bar{Y})^2$	p

Goodness of fit - II

- As in the simple linear model, the goodness of fit can be investigated through the F-test.

$$\mathrm{MSTO} = \frac{\mathrm{SSTO}}{n-1}, \quad \mathrm{MSE} = \frac{\mathrm{SSE}}{n-p-1}, \quad \mathrm{MSR} = \frac{\mathrm{SSR}}{p}$$

$$F = \frac{\mathrm{MSR}}{\mathrm{MSE}}$$

- The F-test is used to test the hypothesis "all $\beta = 0$" against the alternative "at least one $\beta \neq 0$".
- Larger values of the F statistic indicate more evidence for the alternative (the model explains more).

Interpretation

- Estimated intercept gives us the predicted value of Y when all $X = 0$.
- Estimated slope for X_1 give us the change in Y if X_1 changes by one unit, ceteris paribus.
- Estimated slopes $(\hat{\beta}_1, ..., \hat{\beta}_p)$ can thus be interpreted as partial effects, that is β_k gives the change in Y when X_k increases of one unit and the other elements of X remain the same.

Reference

- Kutner M.H., Nachstsheim C.J., Neter J.: Applied Linear Regression Models, 4° edition, McGraw Hill Irwin (2004), available at https://www.academia.edu/32804953/Applied_Linear_Regression_Models_4th_edition