

Master Program in *Data Science and Business Informatics*

# Statistics for Data Science

Lesson 21 - Non-linear, and multiple linear regression

Salvatore Ruggieri

Department of Computer Science

University of Pisa, Italy

[salvatore.ruggieri@unipi.it](mailto:salvatore.ruggieri@unipi.it)

# Simple linear regression model

SIMPLE LINEAR REGRESSION MODEL. In a *simple linear regression model* for a bivariate dataset  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , we assume that  $x_1, x_2, \dots, x_n$  are nonrandom and that  $y_1, y_2, \dots, y_n$  are realizations of random variables  $Y_1, Y_2, \dots, Y_n$  satisfying

$$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \dots, n,$$

where  $U_1, \dots, U_n$  are *independent* random variables with  $E[U_i] = 0$  and  $\text{Var}(U_i) = \sigma^2$ .

- *Regression line*:  $y = \alpha + \beta x$  with *intercept*  $\alpha$  and *slope*  $\beta$
- Least Square Estimators:  $\hat{\alpha}$  and  $\hat{\beta}$  and  $\hat{\sigma}^2$
- Unbiasedness:  $E[\hat{\alpha}] = \alpha$  and  $E[\hat{\beta}] = \beta$  and  $E[\hat{\sigma}^2] = \sigma^2$
- *Standard errors* (estimates of  $\sqrt{\text{Var}(\hat{\alpha})}$  and  $\sqrt{\text{Var}(\hat{\beta})}$ ):

$$se(\hat{\alpha}) = \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}_n^2}{SXX}\right)}$$

$$se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{SXX}}$$

# Standard error of fitted values (prediction $\pm$ standard error)

- For a given  $x_0$ , the estimator  $\hat{Y} = \hat{\alpha} + \hat{\beta}x_0$  has expectation

$$E[\hat{Y}] = E[\hat{\alpha}] + E[\hat{\beta}]x_0 = \alpha + \beta x_0$$

- Hence,  $\hat{Y}$  is unbiased, and  $\hat{y} = \hat{\alpha} + \hat{\beta}x_0$  is the best estimate for the fitted value at  $x_0$
- Variance of  $\hat{Y}$  is:

[See *sdsln.pdf* Chpt. 2]

$$\text{Var}(\hat{Y}) = \sigma^2 \left( \frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX} \right)$$

- The *standard error* of the fitted value is then the estimate:

$$\text{se}(\hat{y}) = \hat{\sigma} \sqrt{\left( \frac{1}{n} + \frac{(\bar{x}_n - x_0)^2}{SXX} \right)}$$

where

$$SXX = \sum_1^n (x_i - \bar{x}_n)^2 \qquad \hat{\sigma}^2 = \frac{1}{n-2} \sum_1^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

- Prediction uncertainty at  $x_0$  is reported as  $\hat{y} \pm \text{se}(\hat{y})$

**See R script**

# Weighted Least Squares and simple polynomial regression

- Weighted Simple Regression

$$S(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 w_i$$

- ▶  $w_i$  is the weight (or importance) of observation  $(x_i, y_i)$
- ▶ For natural number weights, it is the same as replicating instances

- Polynomial Simple Regression

$$S(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta_1 x_i - \beta_2 x_i^2 - \dots - \beta_k x_i^k)^2$$

- ▶  $Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k + U_i$  for  $i = 1, 2, \dots, n$
- ▶ May suffer from collinearity (see next lesson)

**See R script**

# Non-linear simple regression and transformably linear functions

- $Y_i = f(\alpha, \beta, x_i) + U_i$  for  $i = 1, 2, \dots, n$  for a non-linear function  $f()$

$$S(\alpha, \beta) = \sum_{i=1}^n (y_i - f(\alpha, \beta, x_i))^2$$

- $\arg \min_{\alpha, \beta} S(\alpha, \beta)$  may be without a closed form
  - ▶ use numeric search of the minimum (which may fail to find it!), e.g., gradient descent
- Some  $f()$  can be favourably transformed, e.g.,  $f(\alpha, \beta, x_i) = \alpha x_i^\beta$  (recall Power law, Zipf's)
- Solve  $\log Y_i = \log \alpha + \beta \log x_i + U_i$  *[Linearization]*
- Let  $\hat{\alpha}$  and  $\hat{\beta}$  be the LSE estimators. By exponentiation:

$$Y_i = \hat{\alpha} x_i^{\hat{\beta}} e^{U_i}$$

where the error term is a multiplicative factor

**See R script**

# Multiple linear regression

- Multivariate dataset of observations:

$$(x_1^1, x_1^2, \dots, x_1^k, y_1), \dots, (x_n^1, x_n^2, \dots, x_n^k, y_n)$$

- $Y_i = \alpha + \beta_1 x_i^1 + \dots + \beta_k x_i^k + U_i$

- In vector terms:

- ▶  $Y_i = \mathbf{x}_i \cdot \boldsymbol{\beta}^T + U_i$ , where  $\boldsymbol{\beta} = (\alpha, \beta_1, \dots, \beta_k)$  and  $\mathbf{x}_i = (1, x_i^1, \dots, x_i^k)$  the  $i^{\text{th}}$  observation
- ▶  $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta}^T + \mathbf{U}$ , where:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1, x_1^1, x_1^2, \dots, x_1^k \\ 1, x_2^1, x_2^2, \dots, x_2^k \\ \dots \\ 1, x_n^1, x_n^2, \dots, x_n^k \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \dots \\ U_n \end{pmatrix}$$

# Multiple linear regression

- Model:  $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta}^T + \mathbf{U}$
- Ordinary Least Square Estimation (OLS):

$$S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i \cdot \boldsymbol{\beta}^T)^2 = \|\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\beta}^T\|^2 \quad \hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$$

where  $\|(v_1, \dots, v_n)\| = \sqrt{\sum_{i=1}^n v_i^2}$  is the Euclidian norm, and:

$$\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\beta}^T = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} - \begin{pmatrix} 1, x_1^1, x_1^2, \dots, x_1^k \\ 1, x_2^1, x_2^2, \dots, x_2^k \\ \dots \\ 1, x_n^1, x_n^2, \dots, x_n^k \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}$$

- Meaning of  $\beta_i$ : change of  $Y$  due to a unit change in  $x_i$  all the  $x_j$  with  $j \neq i$  unchanged!
- It is a Minimum Variance linear Unbiased Estimator [Gauss-Markov Thm.]

See R script

# Multivariate multiple linear regression

- The multivariate linear model accommodates two or more dependent variables

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^T + \mathbf{U}$$

$$\begin{pmatrix} Y_1^1, \dots, Y_1^m \\ Y_2^1, \dots, Y_2^m \\ \dots \\ Y_n^1, \dots, Y_n^m \end{pmatrix} = \begin{pmatrix} 1, x_1^1, x_1^2, \dots, x_1^k \\ 1, x_2^1, x_2^2, \dots, x_2^k \\ \dots \\ 1, x_n^1, x_n^2, \dots, x_n^k \end{pmatrix} \begin{pmatrix} \alpha^1, \dots, \alpha^m \\ \beta_1^1, \dots, \beta_1^m \\ \dots \\ \beta_k^1, \dots, \beta_k^m \end{pmatrix} + \begin{pmatrix} U_1^1, \dots, U_1^m \\ U_2^1, \dots, U_2^m \\ \dots \\ U_n^1, \dots, U_n^m \end{pmatrix}$$

- $\mathbf{Y}$  is  $n \times m$ :  $n$  observations,  $m$  dependent variables
  - $\mathbf{X}$  is  $n \times (k + 1)$ :  $n$  observations,  $k$  independent variables +1 constants
  - $\boldsymbol{\beta}^T$  is  $(k + 1) \times m$ : parameters for each of the  $m$  dependent variables
  - $\mathbf{U}$  is  $n \times m$ :  $n$  observations,  $m$  error terms
- It is **not** just a collection of  $m$  multiple linear regressions
  - Errors in columns of  $\mathbf{U}$ , e.g.,  $U_1^1, \dots, U_n^1$ , are independent, as in a single multiple linear regression
  - Errors in rows (dependent variables) are allowed to be correlated.
    - E.g., errors of plasma level (e.g.,  $U_1^1$ ) and amitriptyline (e.g.,  $U_1^2$ ) due to usage of drugs
    - Hence, coefficients from the models for the various dependent variables covary!

See R script



# Other variants and generalizations

- Heteroscedastic linear models
  - ▶ Relax the assumption of equal variances  $Var(U_i) = \sigma^2$
- **Generalized least squares**
  - ▶  $U_1, \dots, U_n$  not necessarily independent
- **Hierarchical linear models**
  - ▶ Nested or cluster organization (e.g., Children within classrooms within schools)
  - ▶ See **this intro in R**
- Generalized linear models
  - ▶ We'll see next at Logistic Regression
- **Tobit regression**
  - ▶ Censored dependent variable, e.g., income cannot be negative
- **Truncated regression model**
  - ▶ Dependent variable not available/sampled, e.g., income above a poverty threshold
- **Quantile regression**
  - ▶ Estimate of the median (or other quantiles) instead of the mean, as in regression

# Optional references



Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li (2005)

Applied Linear Statistical Models.

5th edition *McGraw-Hill*