

Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 20 - Linear Regression and Least Squares Estimation

Salvatore Ruggieri

Department of Computer Science

University of Pisa, Italy

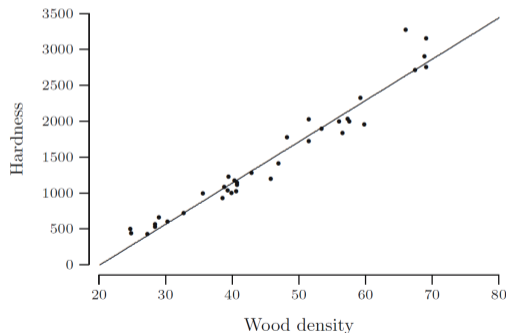
salvatore.ruggieri@unipi.it

Bivariate dataset

- Consider a bivariate dataset

$$(x_1, y_1), \dots, (x_n, y_n)$$

- It can be visualized in a scatter plot



- This suggests a relation $Hardness = \alpha + \beta \cdot Density + random\ fluctuation$

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

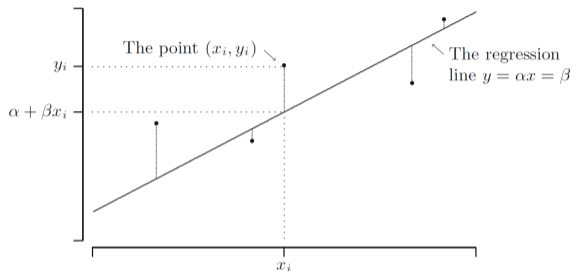
$$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \dots, n,$$

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* α and *slope* β
- x is the *explanatory* (or *independent*) variable, and y the *response* (or *dependent*) variable
- Independence of U_1, \dots, U_n implies independence of Y_1, \dots, Y_n *[propagation of ind.]*
 - ▶ But Y_i 's are not identically distributed, as $E[Y_i] = \alpha + \beta x_i$
- Also, notice the assumption $\text{Var}(Y_i) = \text{Var}(U_i) = \sigma^2$ *[homoscedasticity]*

Estimation of parameters

- How to estimate α and β ? MLE requires to know the distribution of the U_i 's



- $y_i - \alpha - \beta x_i$ is called a *residual* (or the *error*), and it is a realization of U_i
 - ▶ recall that $E[U_i] = 0$ and $Var(U_i) = E[U_i^2] = \sigma^2$
- The method of *Least Squares* prescribes to minimize the sum of squares of residuals:

$$\hat{\alpha}, \hat{\beta} = \arg \min_{\alpha, \beta} S(\alpha, \beta) \quad \text{where } S(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

- ▶ $S(\alpha, \beta)$ also called Sum of Squares of Errors (SSE) or Residual Sum of Squares (RSS)

Least Squares Estimates

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

- Partial derivatives:

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

- Equal to 0 for:

$$n\alpha + \beta \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad \alpha \sum_{i=1}^n x_i + \beta \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

and solving, we get:

$$\hat{\alpha} = \bar{y}_n - \hat{\beta} \bar{x}_n \quad \hat{\beta} = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

- $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ are called the *fitted values*

Ordinary Least Squares (OLS) Estimates

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n \quad \hat{\beta} = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

- Equivalent form of $\hat{\beta}$

[prove it!]

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{SXX} = r_{xy} \frac{s_y}{s_x}$$

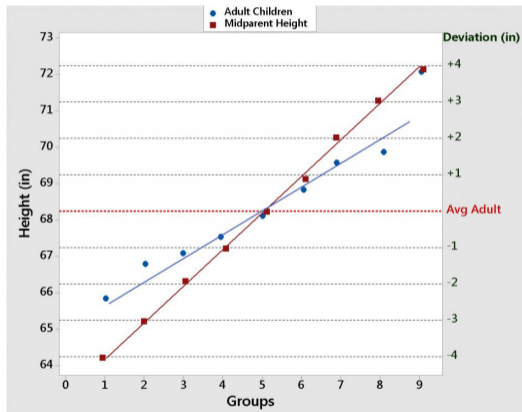
where:

- ▶ $SXX = \sum_{i=1}^n (x_i - \bar{x}_n)^2$
- ▶ $r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$ is the Pearson's correlation coefficient
- ▶ $s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2}$ is the sample standard deviations of x_i 's
- ▶ $s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2}$ is the sample standard deviations of y_i 's
- The line $y = \hat{\alpha} + \hat{\beta}x$ always passes through the *center of gravity* (\bar{x}_n, \bar{y}_n)
 - ▶ Since $\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n$, we have $\hat{\alpha} + \hat{\beta}\bar{x}_n = \bar{y}_n - \hat{\beta}\bar{x}_n + \hat{\beta}\bar{x}_n = \bar{y}_n$

See R script

Why 'regression'?

So, why is it called 'regression' anyway?



“**See Francis Galton** concluded that as heights of the parents deviated from the average height, [...] the heights of the children *regressed* to the average height of an adult.”

Unbiasedness of estimators: $\hat{\beta}$

- Consider the least square estimators:

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n \qquad \hat{\beta} = \frac{\sum_1^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{SXX}$$

where $SXX = \sum_1^n (x_i - \bar{x}_n)^2$. Since $\sum_1^n (x_i - \bar{x}_n) = 0$, we can rewrite $\hat{\beta}$ as:

$$\hat{\beta} = \frac{\sum_1^n (x_i - \bar{x}_n)Y_i - \sum_1^n (x_i - \bar{x}_n)\bar{Y}_n}{SXX} = \frac{\sum_1^n (x_i - \bar{x}_n)Y_i}{SXX} \quad (1)$$

- We have:

$$E[\hat{\beta}] = \frac{\sum_1^n (x_i - \bar{x}_n)E[Y_i]}{SXX} = \frac{\sum_1^n (x_i - \bar{x}_n)(\alpha + \beta x_i)}{SXX} = \frac{\beta \sum_1^n (x_i - \bar{x}_n)x_i}{SXX} = \beta$$

where the last step follows since $\sum_1^n (x_i - \bar{x}_n)x_i = \sum_1^n (x_i - \bar{x}_n)x_i - \sum_1^n (x_i - \bar{x}_n)\bar{x} = SXX$.

- Moreover:

$$\text{Var}(\hat{\beta}) = \frac{\sum_1^n (x_i - \bar{x}_n)^2 \text{Var}(Y_i)}{SXX^2} = \sigma^2 \frac{\sum_1^n (x_i - \bar{x}_n)^2}{SXX^2} = \frac{\sigma^2}{SXX}$$

Unbiasedness of estimators: $\hat{\alpha}$

- Consider the least square estimators:

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n \qquad \hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{SXX}$$

- We have:

$$\begin{aligned} E[\hat{\alpha}] &= E[\bar{Y}_n] - \bar{x}_n E[\hat{\beta}] = \frac{1}{n} \sum_{i=1}^n E[Y_i] - \bar{x}_n \beta \\ &= \frac{1}{n} \sum_{i=1}^n (\alpha + \beta x_i) - \bar{x}_n \beta = \alpha + \bar{x}_n \beta - \bar{x}_n \beta = \alpha \end{aligned}$$

- Moreover:

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

where $\text{Cov}(\bar{Y}_n, \hat{\beta}) = 0$

[prove it or see sdsln.pdf Chpt. 2]

An estimator for σ^2 , and standard errors

- $Var(\hat{\alpha})$ and $Var(\hat{\beta})$ use σ^2 , which is unknown
- We cannot use $\frac{1}{(n-1)} \sum_1^n (Y_i - \bar{Y}_n)^2$ as an estimator of σ^2 , because $E[Y_i]$ is not constant
- An unbiased estimate of σ^2 is:

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

$\hat{\sigma}$ is called the *residual standard error*

- The *standard errors* of the coefficient estimators are defined as the estimates of the standard deviations:

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

See R script

LSE: Relation with MLE

$$Y_i = \alpha + \beta x_i + U_i$$

- In case $U_i \sim N(0, \sigma^2)$, we have $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

- Log-likelihood is

$$\ell(\alpha, \beta) = \sum_{i=1}^n \log \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{y_i - \alpha - \beta x_i}{\sigma} \right)^2} \right) = -n \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

- It turns out that $\arg \max_{\alpha, \beta} \ell(\alpha, \beta) = \hat{\alpha}, \hat{\beta}$ *[same estimators as LSE]*

Total variability = explained variability + unexplained variability

- Total variability in the data. Sum of Squares Total (SST):

$$SST = \sum_1^n (y_i - \bar{y}_n)^2$$

- Variability explained by regression. Sum of Squares of Regression (SSR):

$$SSR = \sum_1^n (\hat{\alpha} + \hat{\beta}x_i - \bar{y}_n)^2$$

- Unexplained variability explained. Sum of Squares of Errors (SSE):

$$SSE = \sum_1^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

- It turns out:

[Prove it!]

$$SST = SSR + SSE$$

- $1 - SSE/SST$ (or SSR/SST) is the fraction of explained variability over total variability

Residuals and R^2

- Residual standard error vs Root Mean Squared Error (RMSE):

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_1^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2} \quad RMSE = \sqrt{\frac{1}{n} \sum_1^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}$$

both measure the variability we cannot explain with the regression model

- Compare $\hat{\sigma}^2$ to the variability of data:

$$\hat{\sigma}_y^2 = \frac{1}{n-1} \sum_1^n (y_i - \bar{y}_n)^2$$

through the *adjusted* R^2 :

$$adjR^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}$$

- $adjR^2$ ranges from 0 (no variability explained) to 1 (all variability explained)

Residuals and R^2

- When taking *un-adjusted* variances::

$$\hat{\sigma}^2 = \frac{1}{n} \sum_1^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = RMSE^2 \quad \hat{\sigma}_y^2 = \frac{1}{n} \sum_1^n (y_i - \bar{y}_n)^2$$

we define the *coefficient of determination* $R^2 = 1 - \hat{\sigma}^2/\hat{\sigma}_y^2$

- Alternative definition based on variance of fitted: $R^2 = \hat{\sigma}_{\hat{y}}^2/\hat{\sigma}_y^2$ where

$$\hat{\sigma}_{\hat{y}}^2 = \frac{1}{n} \sum_1^n (\hat{\alpha} + \hat{\beta}x_i - \bar{\hat{y}}_n)^2 \text{ and } \bar{\hat{y}}_n = \frac{1}{n} \sum_1^n (\hat{\alpha} + \hat{\beta}x_i) = \hat{\alpha} + \hat{\beta}\bar{x}_n = \bar{y}_n$$

and then $\hat{\sigma}_{\hat{y}}^2 = SSR/n$

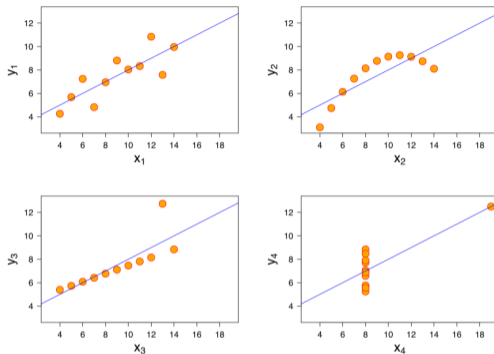
- For simple (one independent r.v.) linear regression:

[Prove it!]

$$R^2 = r_{y\hat{y}} = \frac{\sum_{i=1}^n (y_i - \bar{y}_n) \cdot (\hat{\alpha} + \hat{\beta}x_i - \bar{\hat{y}}_n)}{\sqrt{\sum_{i=1}^n (y_i - \bar{y}_n)^2 \cdot \sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x_i - \bar{\hat{y}}_n)^2}}$$

See R script

Anscombe's quartet



- Regression line $y = 3 + x/2$
 - ▶ Top left: linear regression
 - ▶ Top right: non-linear regression
 - ▶ Bottom left: linear regression with outliers (requires robust regression approaches)
 - ▶ Bottom right: single **high-leverage** point produces correlation
- Look at data graphically before starting to analyze them with a specific technique!