Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 20 - Linear Regression and Least Squares Estimation

Salvatore Ruggieri

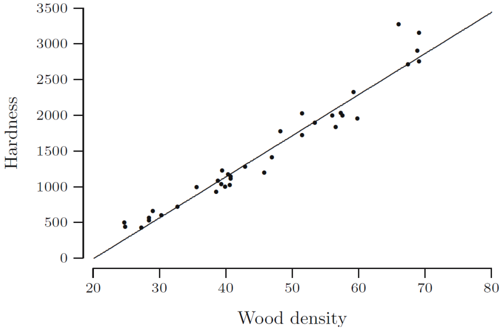
Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it

Bivariate dataset

- Consider a bivariate dataset

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

- 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we assume that $x_1, x_2, \ldots, x_n$ are nonrandom and that $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying

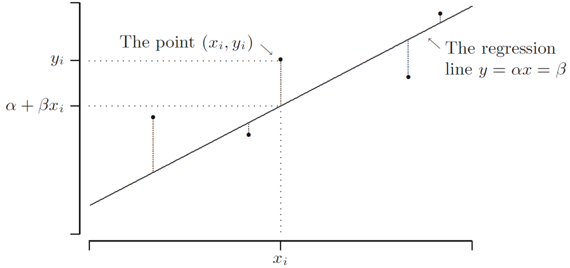
$$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \ldots, n,$$

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

- *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 $U_1, \ldots, U_n$ implies independence of $Y_1, \ldots, Y_n$ *[propagation of indep.]*
  - But Y_i's are not identically distributes, as $E[Y_i] = \alpha + \beta x_i$
- Also, notice the assumption $Var(Y_i) = Var(U_i) = \sigma^2$ *[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 $U_i = Y_i - \alpha - \beta x_i$
  - 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:

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

  - $S(\alpha, \beta)$ 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:

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

are equal to 0 for:

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

and solving, we get:

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

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

Ordinary Least Squares (OLS) Estimates

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

- Equivalent form of $\hat{\beta}$ *[prove it!]*

$$\hat{\beta} = \frac{\sum_1^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{SXX} = r_{xy}\frac{s_y}{s_x}$$

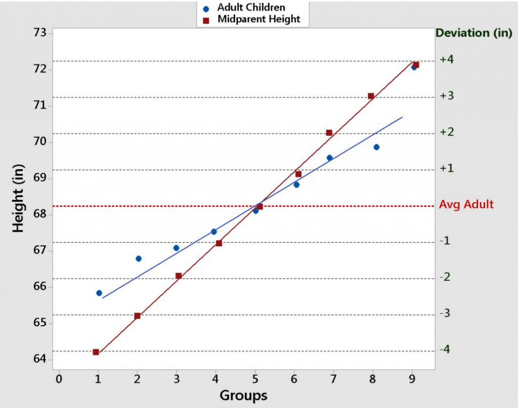
where:
  - $SXX = \sum_1^n (x_i - \bar{x}_n)^2$
  - $r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)\cdot(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2 \cdot \sum_{i=1}^n (y_i - \bar{y}_n)^2}}$ is the Pearson's correlation coefficient
  - $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2}$ 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?



Sir Francis Galton (inventor of standard deviation, regression, and much more)
"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:

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

- 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 $\sum_1^n (x_i - \bar{x}_n) x_i = \sum_1^n (x_i - \bar{x}_n) x_i - \sum_1^n (x_i - \bar{x}_n) \bar{x}_n = SXX$.

- Moreover:

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

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

- Consider the least square estimators:

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

- 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:

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

because $Cov(\bar{Y}_n, \hat{\beta}) = 0$ [*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: *[see Section 22.1 of [T]]*

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

$\hat{\sigma}$ is called the *residual standard error*. A close measure is the Root Mean Squared Error:

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

- 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
  $\ell(\alpha, \beta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma^2}\right)^2}\right) = -n \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$
- It turns out that *arg* $\max_{\alpha,\beta} \ell(\alpha, \beta) = \hat{\alpha}, \hat{\beta}$ *[same estimators as LSE]*
  - *Exercise: prove it!*

Total variability = explained variability + unexplained variability

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

$$SST = \sum_{1}^{n} (y_i - \bar{y}_n)^2$$

- Total variability of the fitted: *explained variability*. Sum of Squares of Regression (SSR):

$$SSR = \sum_{1}^{n} (\hat{\alpha} + \hat{\beta} x_i - \bar{y}_n)^2 = \sum_{1}^{n} (\hat{y}_i - \bar{\hat{y}}_n)^2$$

  because $\bar{\hat{y}}_n = \frac{1}{n} \sum_1^n (\hat{\alpha} + \hat{\beta} x_i) = \hat{\alpha} + \hat{\beta} \hat{x}_n = \bar{y}_n$

- Total variability of residuals: *unexplained variability*. Sum of Squares of Errors (SSE):

$$SSE = \sum_{1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

- It turns out: $SST = SSR + SSE$ *[Prove it!]*
- $1 - SSE/SST$ (or SSR/SST) is the fraction of explained variability over total variability

Residuals and R^2 (fraction of explained variability)

- $1 - SSE/SST$ (or SSR/SST) is the fraction of explained variability over total variability
- When taking sample variances of y's and residuals:

$$\sigma_y^2 = \frac{1}{n-1}\sum_1^n (y_i - \bar{y}_n)^2 = \frac{SST}{n-1} \qquad \sigma_{res}^2 = \frac{1}{n-1}\sum_1^n (y_i - \hat{y}_i)^2 = \frac{SSE}{n-1}$$

  we define the *coefficient of determination* $R^2 = 1 - \sigma_{res}^2/\sigma_y^2$
- Using the sample variance of the fitted:

$$\sigma_{\hat{y}}^2 = \frac{1}{n-1}\sum_1^n (\hat{y}_i - \bar{\hat{y}}_n)^2 = \frac{SSR}{n-1}$$

  we have the alternative (equivalent) definition is $R^2 = \sigma_{\hat{y}}^2/\sigma_y^2$
- For simple (one independent r.v.) linear regression: *[Prove it!]*

$$R^2 = r_{y\hat{y}}^2 = \frac{[\sum_{i=1}^n (y_i - \bar{y}_n) \cdot (\hat{y}_i - \bar{\hat{y}}_n)]^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2 \cdot \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}}_n)^2}$$

Adjusted R^2

- $1 - SSE/SST$ (or SSR/SST) is the fraction of explained variability over total variability
- When taking adjusted sample variances:

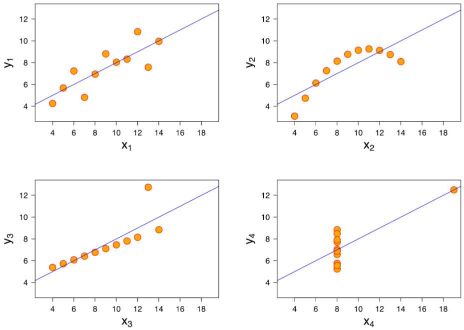
$$\sigma_y^2 = \frac{1}{n-1}\sum_1^n (y_i - \bar{y}_n)^2 = \frac{SST}{n-1} \qquad \hat{\sigma}^2 = \frac{1}{n-2}\sum_1^n (y_i - \hat{y}_i)^2 = \frac{SSE}{n-2}$$

(where $\hat{\sigma}$ is the residual standard error), we define the *adjusted coefficient of determination*:

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

See R script

Anscombe's quartet



- Same regression line $y = 3 + x/2$
  - Top left: linear relation
  - Top right: non-linear relation
  - Bottom left: linear relation 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!

See R script

Optional references

Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li (2005)
Applied Linear Statistical Models.
5th edition *McGraw-Hill*