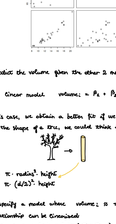
MULTIPLE LINEAR REGRESSION

There are now $p+1$ covariates $x_{i,1}, \dots, x_{i,p}$.
Example: "Area" A district contains data on 31 cherry trees in particular, we have
- diameter (numeric)
- height (float)
- volume

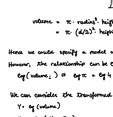
With 2 or more variables we can no longer visualize the relationship with a scatterplot.
We have to use a "matrix of scatterplots" which shows all the PAIRWISE combinations.



The goal is to predict the volume given the other 2 measures

We could use a linear model: $\text{volume}_i = \beta_0 + \beta_1 \text{ diameter}_i + \beta_2 \text{ height}_i + \varepsilon_i$

However, in this case, we obtain a better fit if we consider a transformation of the covariates.
If we think of the shape of a tree, we could think of approximating it to a cylinder.



$$\text{volume} \approx \pi \cdot \text{radius}^2 \cdot \text{height} = \pi \cdot \left(\frac{\text{diameter}}{2}\right)^2 \cdot \text{height}$$

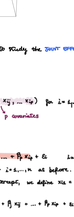
Hence we could specify a model where: $\text{volume}_i \approx \pi \cdot \left(\frac{\text{diameter}_i}{2}\right)^2 \cdot \text{height}_i$ ← not linear

However, the relationship can be linearized:

$$\log(\text{volume}_i) \approx \log\pi + \log 1 + 2\log(\text{diameter}_i) + \log(\text{height}_i)$$

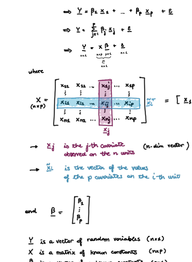
We can consider the transformed variables

- $y_i = \log(\text{volume})$
- $x_{i1} = \log(\text{diameter})$
- $x_{i2} = \log(\text{height})$



With 2 or more covariates we can not see the "pure" effect that one of them has on y, but only the univariate (marginal) effect of a covariate at a time.

Only in the case of two covariates we can think on the joint effect using a 3D representation



The goal of the multiple is to study the JOINT EFFECT of the covariates on y.

MODEL SPECIFICATION

We now observe $(y_i, x_{i1}, x_{i2}, \dots, x_{ip})$ for $i = 1, \dots, n$

response variable; p covariates

Define the model

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \dots, n \end{aligned}$$

with $\varepsilon_i \sim N(0, \sigma^2)$ iid for $i = 1, \dots, n$ as before.

If we want to include the intercept, we define $x_{i0} = 1$ for all $i = 1, \dots, n$ (constant variable) and we obtain the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \dots, n$$

NOTATION:

$$\begin{bmatrix} Y_1 = \beta_0 x_{11} + \beta_1 x_{12} + \dots + \beta_p x_{1p} + \varepsilon_1 \\ \vdots \\ Y_i = \beta_0 x_{i1} + \beta_1 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i \\ \vdots \\ Y_n = \beta_0 x_{n1} + \beta_1 x_{n2} + \dots + \beta_p x_{np} + \varepsilon_n \end{bmatrix} \Rightarrow \underbrace{\begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}}_{\underline{Y}} = \underbrace{\begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{i1} & \dots & x_{ip} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}}_{X} \cdot \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\underline{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\underline{\varepsilon}} \quad \text{for } i = 1, \dots, p$$

$$\Rightarrow \underline{Y} = \beta_0 \underline{x}_0 + \dots + \beta_p \underline{x}_p + \underline{\varepsilon}$$

$$\Rightarrow \underline{Y} = \sum_{j=1}^{p} \beta_j \underline{x}_j + \underline{\varepsilon}$$

$$\Rightarrow \underline{Y} = \underset{(n \times p)}{X} \cdot \underset{(p \times 1)}{\underline{\beta}} + \underset{(n \times 1)}{\underline{\varepsilon}}$$

where

$$\underset{(n \times p)}{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \quad \underline{x}_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix} = \begin{bmatrix} \underline{x}_1 & \underline{x}_2 & \dots & \underline{x}_j & \dots & \underline{x}_p \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$\Rightarrow$ $\underline{x}_j$ is the j^{th} covariate observed on the n units (n-dim vector)

$\Rightarrow$ x_i^T is the vector of the values of the p covariates on the i^{th} unit (p-dim vector)

and $\underline{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$

$\underline{Y}$ is a vector of random variables (nx1)
X is a matrix of known constants (nxp)
$\underline{\beta}$ is a vector of unknown constants (px1)
$\underline{\varepsilon}$ is a vector of random variables (nx1)

The assumptions don't change (they are just adjusted for the general case)

1. normality, homoscedasticity, corr = 0 $\Rightarrow$ $\underline{\varepsilon} \overset{iid}{\sim} N(\underline{0}, \sigma^2 I)$ $\quad i = 1, \dots, n$
2. linearity: $\mu_i = \beta_0 x_{i1} + \dots + \beta_p x_{ip}$
3. absence of multicollinearity of the $\underline{x}_j$: the covariates must be linearly independent

let's analyze the 3 hypothesis:

3. ABSENCE OF MULTICOLLINEARITY

What is the meaning of this hypothesis on $x_1, \dots, x_p$ (i.e. on the matrix X)?
Intuitively, it means that each covariate $\underline{x}_j$ should have an individual contribution for predicting Y
→ the information contained in $\underline{x}_j$ can not be derived from the other variables.
Examples of multicollinearity:
- two variables is expressed with two measurement units (cm/m)
- one variable is a linear combination of the others
  (e.g. x_1 = total years of education; x_2 = years of pre-university education;
  x_3 = years of post-university education; $\Rightarrow$ $x_1 = x_2 + x_3$)

What happens when this hypothesis is not satisfied ?
Assume that $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p$ are linearly dependent:
by definition, it means that there are p scalars $a_1, \dots, a_p$, not all zero, such that
$a_1 \underline{x}_1 + a_2 \underline{x}_2 + \dots + a_p \underline{x}_p = \underline{0}$
This means that I can write the j^{th} variable as:

$$\underline{x}_j = -\frac{a_1}{a_j}\underline{x}_1 - \dots - \frac{a_{j-1}}{a_j}\underline{x}_{j-1} - \frac{a_{j+1}}{a_j}\underline{x}_{j+1} - \dots - \frac{a_p}{a_j}\underline{x}_p \qquad (*)$$

Looking now at our linear model:

$$\underline{Y} = \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \dots + \beta_j \underline{x}_j + \beta_{j+1} \underline{x}_{j+1} + \dots + \beta_p \underline{x}_p + \underline{\varepsilon}$$

We can substitute $\underline{x}_j$ with $(*)$, obtaining:

$$\begin{aligned} \underline{Y} &= \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \dots + \beta_j \underline{x}_j + \beta_{j+1} \underline{x}_{j+1} + \dots + \beta_p \underline{x}_p + \underline{\varepsilon} \\ &= \underbrace{\left(\beta_1 - \beta_j \tfrac{a_1}{a_j}\right)}_{\beta_1^*} \underline{x}_1 + \dots + \underbrace{\left(\beta_{j-1} - \beta_j \tfrac{a_{j-1}}{a_j}\right)}_{\beta_{j-1}^*} \underline{x}_{j-1} + \underbrace{\left(\beta_{j+1} - \beta_j \tfrac{a_{j+1}}{a_j}\right)}_{\beta_j^*} \underline{x}_{j+1} + \dots + \underbrace{\left(\beta_p - \beta_j \tfrac{a_p}{a_j}\right)}_{\beta_{p-1}^*} \underline{x}_p + \underline{\varepsilon} \end{aligned}$$

We have expressed the same model using only $p-1$ variables.
Hence we need to require that the covariates are linearly independent $\Rightarrow$ $\text{rank}(X) = p$

REMARK: p is the number of columns of X
If we include the INTERCEPT in the model $\Rightarrow$ $\underline{1}_n + \underline{x}_1 + \dots + \underline{x}_p$ constant vector of ones
Hence, to avoid collinearity, I can not have another $\underline{x}_j$ that is constant for all $i = 1, \dots, n$
$\Rightarrow$ we can not have $\text{var}(\underline{x}_j) = 0$ for $j = 1, \dots, p$
This is the assumption we have in the graph (we included the intercept)

2. LINEARITY $\quad \mu_i = \sum_{j=1}^{p} \beta_j x_{ij} = x_i^T \underline{\beta}$

LINEAR IN THE PARAMETERS $\underline{\beta}$

3. DISTRIBUTION: normality, homoscedasticity, uncorrelation

$$\underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \text{random} \quad \text{of the errors}$$

- expectation:
$\mathbb{E}[\underline{\varepsilon}] = \underline{0}$ n-dimensional vector of zeros
- variance:

$$\begin{aligned} \text{var}(\underline{\varepsilon}) &= \mathbb{E}\left[(\underline{\varepsilon} - \mathbb{E}[\underline{\varepsilon}])(\underline{\varepsilon} - \mathbb{E}[\underline{\varepsilon}])^T\right] \\ &= \mathbb{E}[\underline{\varepsilon}\underline{\varepsilon}^T] \qquad \text{what is this quantity?} \\ &= \begin{bmatrix} \mathbb{E}[\varepsilon_1^2] & \mathbb{E}[\varepsilon_1 \varepsilon_2] & \dots & \mathbb{E}[\varepsilon_1 \varepsilon_n] \\ \mathbb{E}[\varepsilon_2 \varepsilon_1] & \mathbb{E}[\varepsilon_2^2] & & \vdots \\ \vdots & & \ddots & \\ \mathbb{E}[\varepsilon_n \varepsilon_1] & \dots & & \mathbb{E}[\varepsilon_n^2] \end{bmatrix} \qquad \underline{\varepsilon}\underline{\varepsilon}^T = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \cdot \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{bmatrix} = \begin{bmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \dots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & & \vdots \\ \vdots & & \ddots & \\ \varepsilon_n\varepsilon_1 & \dots & & \varepsilon_n^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & \sigma^2 \end{bmatrix} \qquad \text{since } \begin{cases} \mathbb{E}[\varepsilon_i \varepsilon_k] = 0 & \text{for } i \neq k \\ \mathbb{E}[\varepsilon_i^2] = \sigma^2 & \text{for } i = 1, \dots, n \end{cases} \\ &= \sigma^2 I_n \qquad \begin{matrix} \text{(non) scalar, diagonal elements} = \sigma^2 \\ \text{off-diagonal elements} = 0 \end{matrix} \end{aligned}$$

Hence $\underline{\varepsilon} \overset{iid}{\sim} N(\underline{0}, \sigma^2) \quad i = 1, \dots, n$ $\Rightarrow$ $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I_n)$

consequences for the response variable:
$\mathbb{E}[\underline{Y}] = \mathbb{E}[X\underline{\beta} + \underline{\varepsilon}] = X\underline{\beta}$
$\text{var}(\underline{Y}) = \text{var}(X\underline{\beta} + \underline{\varepsilon}) = \text{var}(\underline{\varepsilon}) = \sigma^2 I_n$
Finally, the normality of $\underline{\varepsilon}$ implies the normality of $\underline{Y}$ $\Rightarrow$ $\underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I_n)$

- INTERPRETATION OF THE COEFFICIENTS $\quad \beta_1, \dots, \beta_p$

We have that β_j is the slope (linear model: $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$, β_j is the expected change in Y_i (i.e., the change in $\mu_i = \mathbb{E}[Y_i]$) when we increase x_{ij} by one unit.
(or, equivalently, the expected difference in Y when we consider two individuals i and k which differ in x_j by 1 unit: $\quad \beta_j = \mathbb{E}[Y_k] - \mathbb{E}[Y_i]$, when $x_{ij} = x_{kj}$ and $x_{ik} = x_{kl} + 1$)

How do we interpret β_j, $j = 1, \dots, p$, in the case of multiple linear regression?

$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$

Let's consider the mean of Y of two units i and k, $\mathbb{E}[Y_i] = \mu_i$ and $\mathbb{E}[Y_k] = \mu_k$.
Assume that the values of the j^{th} covariate of these individuals are $x_{ij} = x_k$ and $x_{kj} = x_k + 1$
while the other covariates are all equal: $x_{i1} = x_{k1}$, $x_{i2} = x_{k2}$, ..., $x_{i,j-1} = x_{k,j-1}$,
$x_{i,j+1} = x_{k,j+1}$, ..., $x_{ip} = x_{kp}$

Let's now compare their means:

$$\begin{aligned} \mu_i &= \beta_0 + \beta_1 x_{i1} + \dots + \beta_j x_{ij} + \dots + \beta_p x_{ip} \qquad \text{mean of individual } i \\ &= \beta_0 + \beta_1 x_{i1} + \dots + \beta_j x_k + \dots + \beta_p x_{ip} \\ \mu_k &= \beta_0 + \beta_1 x_{k1} + \dots + \beta_j (x_k + 1) + \dots + \beta_p x_{kp} \qquad \text{mean of individual } k \\ &= \beta_0 + \beta_1 x_{k1} + \dots + \beta_j x_k + \beta_j + \dots + \beta_p x_{kp} \end{aligned}$$

If we study the difference in their means:

$$\Rightarrow \mu_k - \mu_i = \beta_j$$

β_j now represents the expected change in Y_i (i.e., the change in μ_i), when we increase x_j by one unit, while keeping all other covariates fixed.