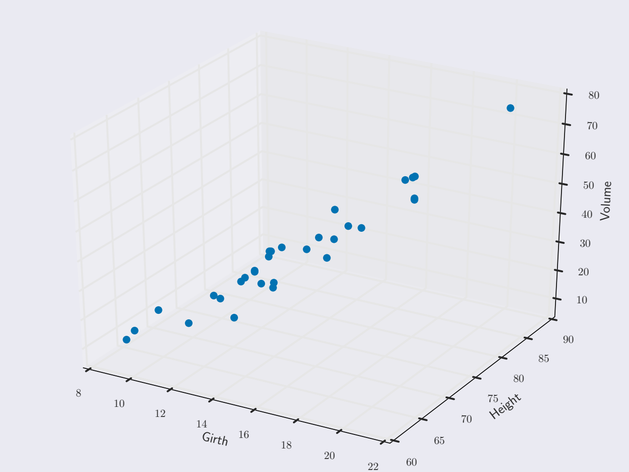
SD-TSIA204
Statistics : linear models

Ekhine Irurozki
Télécom Paris

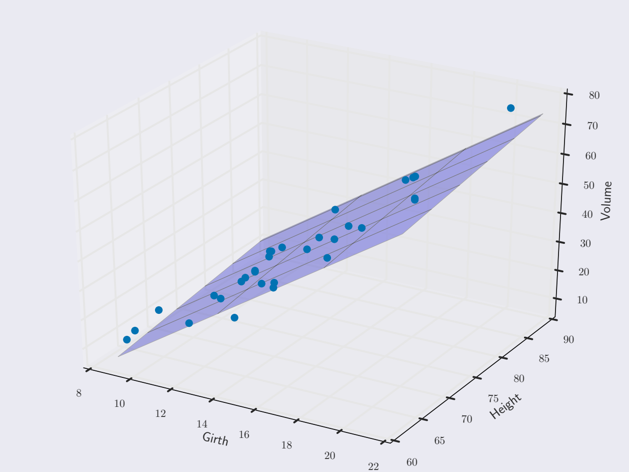
Toward multivariate models

Tree volume as a function of height / girth (🇫🇷 : *circonférence*)



Toward multivariate models

Tree volume as a function of height / girth (🇫🇷 : *circonférence*)



Python commands

```
from matplotlib.mplot3d import Axes3D
# Load data
url = 'http://vincentarelbundock.github.io/
       Rdatasets/csv/datasets/trees.csv'
dat3 = pd.read_csv(url)
# Fit regression model
X = dat3[['Girth', 'Height']]
X = sm.add_constant(X)
y = dat3['Volume']
results = sm.OLS(y, X).fit().params
XX = np.arange(8, 22, 0.5)
YY = np.arange(64, 90, 0.5)
xx, yy = np.meshgrid(XX, YY)
zz = results[0] + results[1]*xx + results[2]*yy
fig = plt.figure()
ax = Axes3D(fig)
ax.plot(X['Girth'],X['Height'],y,'o')
ax.plot_wireframe(xx, yy, zz, rstride=10, cstride=10)
plt.show()
```

results output `const:-57.98, Girth: 4.70, Height: 0.33`

Model

One observes p features $(\mathbf{x}_1, \dots, \mathbf{x}_p)$. Model in dimension p

$$y_i = \theta_0^\star + \sum_{j=1}^{p} \theta_j^\star x_{i,j} + \varepsilon_i$$

$$\varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ pour } i = 1, \dots, n$$

$$\mathbb{E}(\varepsilon) = 0$$

<u>Rem</u>:we assume (frequentist point of view) there exists a "true" parameter $\boldsymbol{\theta}^\star = (\theta_0^\star, \dots, \theta_p^\star)^\top \in \mathbb{R}^{p+1}$

Dimension p

Matrix model

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \theta_0^\star \\ \vdots \\ \theta_p^\star \end{pmatrix}}_{\boldsymbol{\theta}^\star} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

Equivalently : $\boxed{\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}}$

Column notation : $X = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_p)$ with $\mathbf{x}_0 = \mathbf{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

Line notation : $X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} = (x_1, \dots, x_n)^\top$

Rem:often $\mathbf{x}_0$ will be omitted by simplicity, *e.g.*,center $\mathbf{y}$ first

Vocabulary

$$\boxed{\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}}$$

- $\mathbf{y} \in \mathbb{R}^n$: observations vector
- $X \in \mathbb{R}^{n\times(p+1)}$: **design** matrix (with features as columns)
- $\boldsymbol{\theta}^\star \in \mathbb{R}^{p+1}$: (unknown) **true** parameter to be estimated
- $\boldsymbol{\varepsilon} \in \mathbb{R}^n$: noise vector

(Ordinary) Least squares

<u>A</u> least square estimator is **<u>any</u>** solution of the following problem :

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min} \left(\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \right)$$

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left[y_i - \left(\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j} \right) \right]^2$$

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left[y_i - \langle x_i, \boldsymbol{\theta} \rangle \right]^2$$

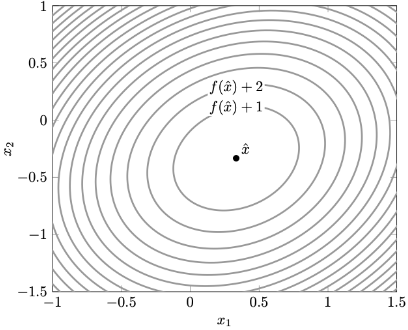
- Does the solution exist ? A solution always exists, as we are minimizing a coercive continuous function (**coercive** : $\lim_{\|x\| \to +\infty} f(x) = +\infty$)
- Is the solution unique ? not guaranteed

The system of equations $\mathbf{y} = X\boldsymbol{\theta}^\star$

General systems of equations $Ax = b$ for $a \in \mathbb{R}^{n\times(p+1)}$ and $p > n$ has no solution

$$A = \begin{bmatrix} 2,0 \\ -1,1 \\ 0,2 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

OLS is to minimize $\|Ax - b\|_2^2$



Row / column interpretation

- Let $\tilde{x}_1^\top, \ldots, \tilde{x}_{p+1}^\top$ be the rows of X. The residuals are $r_i = \tilde{x}_i \boldsymbol{\theta} - y_i$ and the OLS is equivalent to minimizing the sum of squares residuals
- Let $x_1, \ldots, x_{p+1}$ be the columns of X. Then $\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 = \|(\theta_0 x_0, \ldots, \theta_p x_p) - \mathbf{y}\|_2^2$, so OLS is to find a linear combination of columns of X that is closest to $\mathbf{y}$.

Vocabulary (and abuse of terms)

We call **Gram matrix** the matrix

$$X^\top X$$

whose general term is $[X^\top X]_{i,j} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
Rem:$X^\top X$ is often referred to as the feature correlation matrix (true for standardized columns)

Rem: when columns are scaled such that $\forall j \in [\![0, p]\!], \|\mathbf{x}_j\|^2 = n$, the Gramian diagonal is $(n, \ldots, n)$

The vector $X^\top \mathbf{y} = \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{pmatrix}$ represents the correlation

between the observations and the features

Hilbert projection theorem (HPT)

Let $C \subset \mathbb{R}^d, Y \in \mathbb{R}^d$. Let $\hat{z} = \arg\min_{z \in C} \|Y - z\|_2^2$. Then $\hat{z}$ always exists and is given by

$$\boxed{< Y - \hat{z}, z >= 0 \qquad \forall z \in C}$$

Hilbert projection theorem (HPT) and application to OLS

HPT : Let $C \subset \mathbb{R}^d, Y \in \mathbb{R}^d$. Let $\hat{z} = \arg\min_{z \in C} \|Y - z\|_2^2$. Then $\hat{z}$ always exists and is given by

$$\boxed{< Y - \hat{z}, z >= 0 \qquad \forall z \in C}$$

Recall the OLS, $\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \left(\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2\right)$. Define $C = span(X)$ (i.e., the columns of the design matrix X), we redefine the OLS problem as $\hat{Z} \in \arg\min_{Z \in C} \left(\|\mathbf{y} - Z\|_2^2\right)$ and use the characterization of $\hat{Z}$ of the HPT.

$$\begin{aligned}
\langle \hat{Z} - \mathbf{y}, Z \rangle &= 0 \quad \forall Z \\
(\hat{Z} - \mathbf{y})^\top, Z &= 0 \quad \forall Z \\
(\hat{Z} - \mathbf{y})^\top, X\boldsymbol{\theta} &= 0 \quad \forall \boldsymbol{\theta} \\
(\hat{Z} - \mathbf{y})^\top, X &= 0 \\
X^\top(\hat{Z} - \mathbf{y}) &= 0 \\
X^\top(X\hat{\boldsymbol{\theta}} - \mathbf{y}) &= 0
\end{aligned}$$

OLS normal equations

The solution to the OLS problem is given by the solution to the normal equation

Normal equation : $$\boxed{X^\top X\hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$$

As a consequence,

- a solutions always exists.
- its unique if the solution to the normal equations is unique

Hilbert projection theorem

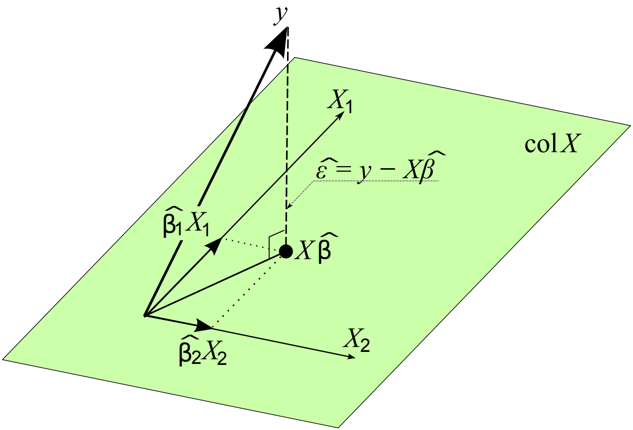


FIGURE – Souce : Wikipedia

Least squares and uniqueness

Let $\hat{\boldsymbol{\theta}}$ be a solution of $\boxed{X^\top X \hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$

Non uniqueness : happens for non trivial kernel,*i.e.*,when $\mathrm{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$

Assume $\boldsymbol{\theta}_K \in \mathrm{Ker}(X)$ with $\boldsymbol{\theta}_K \neq 0$, then

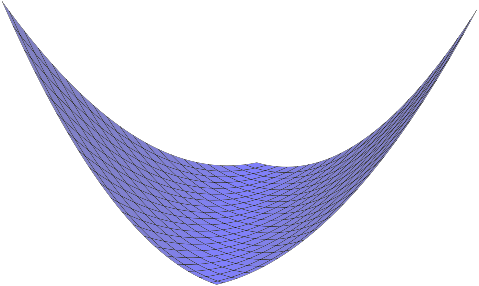
$$
\begin{aligned}
X(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) =& X\hat{\boldsymbol{\theta}} \\
\text{and then} \quad (X^\top X)(\hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_K) =& X^\top \mathbf{y}
\end{aligned}
$$

<u>Conclusion</u> : the set of least squares solutions is an affine sub-space

$$\boxed{\hat{\boldsymbol{\theta}} + \mathrm{Ker}(X)}$$

Optimization in $\mathbb{R}^d$

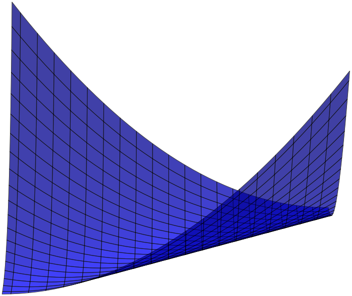
Convex case, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

Optimization in $\mathbb{R}^d$

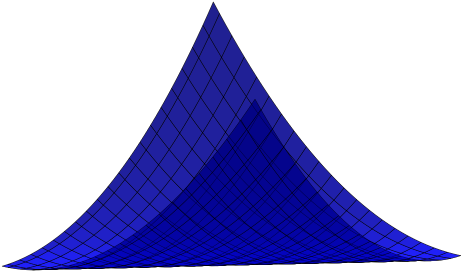
Convex case, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

Optimization in $\mathbb{R}^d$

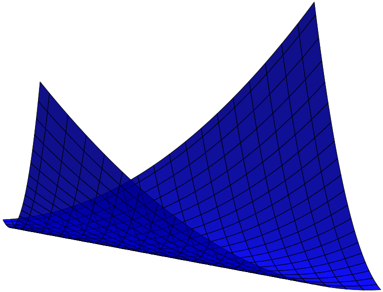
Convex case, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

Optimization in $\mathbb{R}^d$

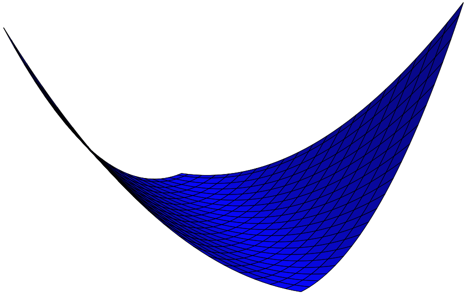
Convex case, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

Optimization in $\mathbb{R}^d$

Convex case, $f(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$, where the set of minimizers is non-unique :



Rem: here the set of minimizers is a line

Non uniqueness : single feature case

Reminder : $$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

If $\mathrm{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^2 : X\boldsymbol{\theta} = 0\} \neq \{0\}$ there exists $(\theta_0, \theta_1) \neq (0, 0)$:

$$\begin{cases} \theta_0 + \theta_1 x_1 & = 0 \\ \vdots \qquad \vdots & = \vdots \\ \theta_0 + \theta_1 x_n & = 0 \end{cases} \qquad (\star)$$

1. If $\theta_1 = 0$: $(\star) \Rightarrow \theta_0 = 0$, so $(\theta_0, \theta_1) = (0, 0)$, **contradiction**
2. If $\theta_1 \neq 0$:
   2.1 If $\forall i, x_i = 0$ then $X = (\mathbf{1}_n, 0)$ and $\theta_0 = 0$
   2.2 Otherwise there exists $x_{i_0} \neq 0$ and $\forall i, x_i = -\theta_0/\theta_1 = x_{i_0}$, *i.e.,*$X = [\mathbf{1}_n \quad x_{i_0} \cdot \mathbf{1}_n]$

Interpretation : $\mathbf{x}_1 \propto \mathbf{1}_n$, *i.e.,*$\mathbf{x}_1$ is constant

Interpretation for multivariate cases

Reminder : we write $X = (\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p)$, the features being column-wise (each are of length n)

The property $\mathrm{Ker}(X) = \{\boldsymbol{\theta} \in \mathbb{R}^{p+1} : X\boldsymbol{\theta} = 0\} \neq \{0\}$ means that there exists a linear dependence between the features $\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p$,

Reformulation : $\exists \boldsymbol{\theta} = (\theta_0, \ldots, \theta_p)^\top \in \mathbb{R}^{p+1} \backslash \{0\}$ s.t.

$$\theta_0 \mathbf{1}_n + \sum_{j=1}^{p} \theta_j \mathbf{x}_j = 0$$

Algebra reminder

Definition

Rank of a matrix : $\quad \mathrm{rank}(X) = \dim(\mathrm{Span}(\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p))$;
$\mathrm{Span}(\cdot)$: the space generated by $\cdot$

Property : $\mathrm{rank}(X) = \mathrm{rank}(X^\top)$

Rank–nullity theorem

$$\mathrm{rank}(X) + \dim(\mathrm{Ker}(X)) = p + 1$$
$$\mathrm{rank}(X^\top) + \dim(\mathrm{Ker}(X^\top)) = n$$

Rem: $\boxed{\mathrm{rank}(X) \leqslant \min(n, p+1)}$

See Golub and Van Loan (1996) for details

Exercise: $\mathrm{Ker}(X) = \mathrm{Ker}(X^\top X)$

Algebra reminder (continued)

Matrix inversion

A square matrix $A \in \mathbb{R}^{m \times m}$ is invertible

- if and only if its kernel is trivial : $\mathrm{Ker}(A) = \{0\}$
- if and only if it is full rank $\mathrm{rank}(A) = m$

Exercise: Show that $\mathrm{Ker}(A) = \{0\}$ is equivalent to $A^\top A$ invertible

---

Closed-form solution for least squares

Closed-form solution for full rank matrix

If X is full (column) rank (*i.e.*,if $X^\top X$ is non-singular) then

$$\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

Rem: recover the empirical mean if $X = \mathbf{1}_n$: $\hat{\boldsymbol{\theta}} = \dfrac{\langle \mathbf{1}_n, \mathbf{y} \rangle}{\langle \mathbf{1}_n, \mathbf{1}_n \rangle} = \bar{y}_n$

Rem: for a single feature $X = \mathbf{x} = (x_1, \dots, x_n)^\top$: $\hat{\boldsymbol{\theta}} = \langle \dfrac{\mathbf{x}}{\|\mathbf{x}\|^2}, \mathbf{y} \rangle$

Beware : in practice **avoid** inverting the matrix $X^\top X$:

- this is numerically time consuming
- the matrix $X^\top X$ might be big if "$p \gg n$", *e.g.*,in biology n patients (≈ 100), p genes (≈ 50000)

Exercise: recover formula for 1D case with intercept

Prediction

Definition

$$\textbf{Prediction vector :} \qquad \hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}}$$

Rem: $\hat{\mathbf{y}}$ depends linearly on the observation vector $\mathbf{y}$

Reminder : an **orthogonal projector** is a matrix H such that

1. H is symmetric : $H^\top = H$
2. H is idempotent : $H^2 = H$

Proposition

Writing H_X the orthogonal projector onto the space span by the columns of X, one gets $\hat{\mathbf{y}} = H_X\mathbf{y}$

Rem: if X is full (column) rank, then $H_X = X(X^\top X)^{-1}X^\top$ is called the **hat matrix**

Prediction (continued)

If a new observation $x_{n+1} = (x_{n+1,1}, \dots, x_{n+1,p})$ is provided, the associated prediction is :

$$\hat{y}_{n+1} = \langle \hat{\boldsymbol{\theta}}, (1, x_{n+1,1}, \dots, x_{n+1,p})^\top \rangle$$
$$\hat{y}_{n+1} = \hat{\theta}_0 + \sum_{j=1}^{p} \hat{\theta}_j x_{n+1,j}$$

Rem: the normal equation ensures **equi-correlation** between observations and features :

$$\begin{aligned}(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y} &\Leftrightarrow X^\top \hat{\mathbf{y}} = X^\top \mathbf{y} \\ &\Leftrightarrow \begin{pmatrix} \langle \mathbf{x}_0, \hat{\mathbf{y}} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \hat{\mathbf{y}} \rangle \end{pmatrix} = \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{pmatrix}\end{aligned}$$

Let $P = \begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \in \mathbb{R}^{n\times n}$.

1. Check that P is an orthogonal projection matrix
2. Determine $\mathrm{Im}(P)$,the range of P
3. For $\mathbf{x} = (x_1, \ldots, x_n)^\top$, $\overline{x}_n$ is the empirical mean and $\sigma_{\mathbf{x}}$ is the standard deviation :

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \sigma_{\mathbf{x}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}.$$

   Show that $\sigma_{\mathbf{x}} = \|(\mathrm{Id}_n - P)\mathbf{x}\|/\sqrt{n}$.

Residuals and normal equation

Definition

Residual(s) : $\quad \mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\theta}} = (\mathrm{Id}_n - H_X)\mathbf{y}$

Reminder :

Normal Equation : $\boxed{(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$

Thanks to the residual definition, the later yields :

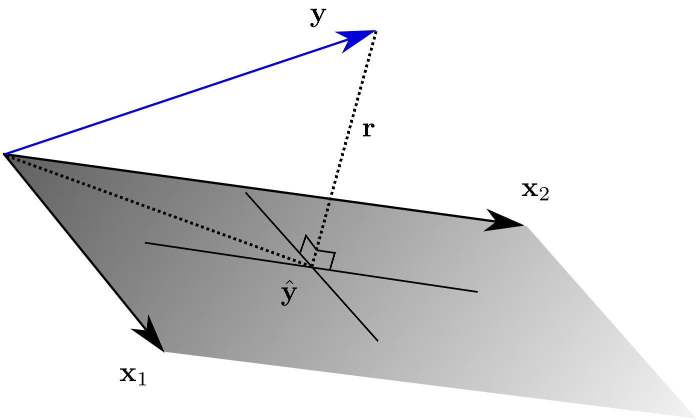
$$X^\top (X\hat{\boldsymbol{\theta}} - \mathbf{y}) = 0 \Leftrightarrow X^\top \mathbf{r} = 0 \Leftrightarrow \mathbf{r}^\top X = 0$$

With $X = (\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p)$, this can be rewritten

$$\forall j = 1, \ldots, p : \langle \mathbf{r}, \mathbf{x}_j \rangle = 0 \text{ and } \overline{r}_n = 0$$

Interpretation : residuals are orthogonal to features

Visualization : predictors and residuals $(p = 2)$



References I