Model pomiaru

Measurements of the speed of light in air, made between 5th June and 2nd July, 1879. The data consists of five experiments, each consisting of 20 consecutive runs. The response is the speed of light in km/s, less 299000. The currently accepted value, on this scale of measurement, is 734.5.

Speed	Run	Expt	Speed	Run	Expt	Speed	Run	Expt	Speed	Run	Expt	Speed	Run	Expt
850	1	1	850	6	1	1000	11	1	810	16	1	960	1	2
740	2	1	950	7	1	980	12	1	1000	17	1	940	2	2
900	3	1	980	8	1	930	13	1	1000	18	1	960	3	2
1070	4	1	980	9	1	650	14	1	960	19	1	940	4	2
930	5	1	880	10	1	760	15	1	960	20	1	880	5	2

Analiza wariancji

Analiza wariancji jednoczynnikowej

Dataset comes from a study of blood coagulation times [sec]. 24 animals were randomly assigned to four different diets and the samples were taken in a random order.

coag	diet	coag	diet	coag	diet	coag	diet
62	A	63	B	68	C	56	D
60	A	67	B	66	C	62	D
63	A	71	B	71	C	60	D
59	A	64	B	67	C	61	D
63	B	65	B	68	C	63	D
		66	B	68	C	64	D
						63	D
						59	D

Regresja

```
> data("gala")
```

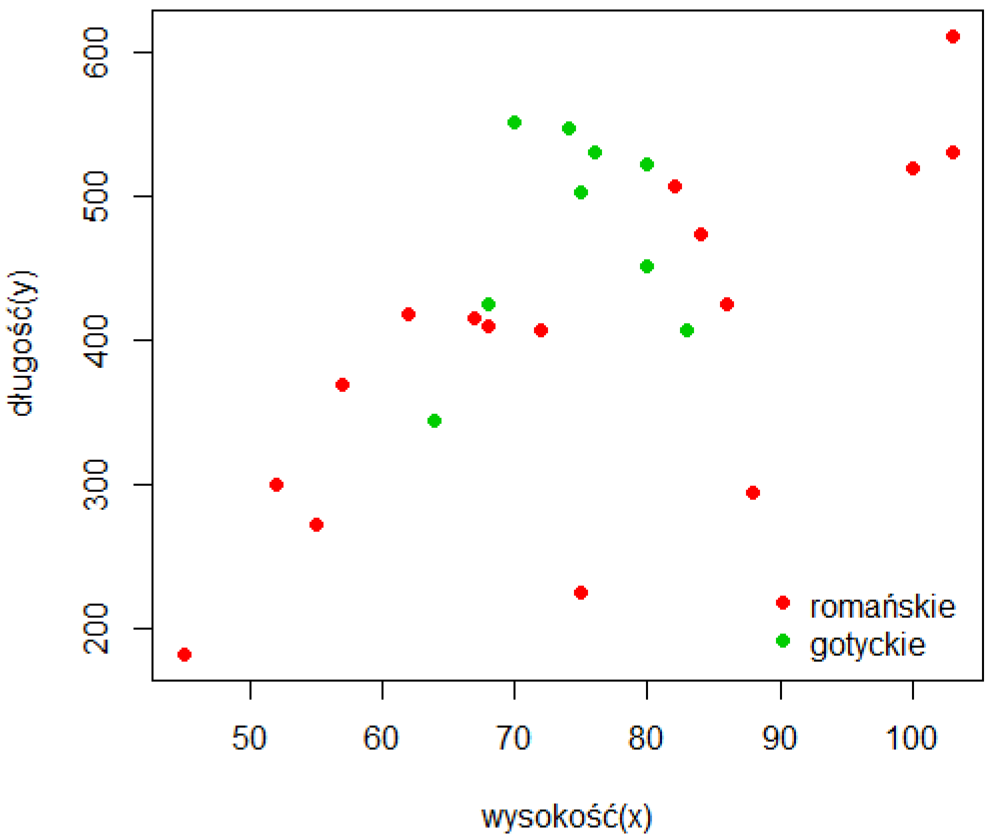
There are 30 Galapagos islands and 7 variables in the dataset. The relationship between the number of plant species and several geographic variables is of interest.

```
> head(gala)
             Species Endemics  Area Elevation Nearest Scruz Adjacent
Baltra            58       23 25.09       346     0.6   0.6     1.84
Bartolome         31       21  1.24       109     0.6  26.3   572.33
Caldwell           3        3  0.21       114     2.8  58.7     0.78
Champion          25        9  0.10        46     1.9  47.4     0.18
Coamano            2        1  0.05        77     1.9   1.9   903.82
Daphne.Major      18       11  0.34       119     8.0   8.0     1.84
```

Species The number of species ~~of tortoise~~ found on the island
Endemics The number of endemic species
Elevation The highest elevation of the island (m)
Nearest The distance from the nearest island (km)
Scruz The distance from Santa Cruz island (km)
Adjacent The area of the adjacent island (km2)

g1 <- lm(species~elevation,data=gala.eaa) – ***regresja jednej zmiennej***

g3 <- lm(species~area+elevation+adjacent,data=gala.eaa) – ***regresja wielu zmiennych***



Analiza kowariancji

	style	x	y		style	x	y
Durham	r	75	502	York	g	100	519
Canterbury	r	80	522	Bath	g	75	225
Gloucester	r	68	425	Bristol	g	52	300
Hereford	r	64	344	Chichester	g	62	418

Model logistyczny

Number of damage incidents out of 6 possible

Presidential Commission on the Space Shuttle Challenger Accident, Vol. 1, 1986: 129-131.

temp	damage	temp	damage	temp	damage	temp	damage
53	5	68	0	70	0	76	0
57	1	69	0	72	0	76	0
58	1	70	1	73	0	78	0
63	1	70	0	75	0	79	0
66	0	70	1	75	1	81	0
67	0	67	0	67	0		

Wspólna reprezentacja

TODO

1.2.1 THE LINEAR REGRESSION MODEL

The data consist of n sets of observations $\{x_{1i}, x_{2i}, \ldots, x_{pi}, y_i\}$, which represent a random sample from a larger population. It is assumed that these observations satisfy a linear relationship,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \qquad (1.1)$$

where the β coefficients are unknown parameters, and the ε_i are random error terms. By a *linear* model, it is meant that the model is linear in the *parameters*; a quadratic model,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

*

The errors are normally distributed. This is needed if we want to construct any confidence or prediction intervals, or hypothesis tests, which we usually do. If this assumption is violated, hypothesis tests and confidence and prediction intervals can be very misleading.

$$(E(\varepsilon_i) = 0 \text{ for all } i).$$

$$(V(\varepsilon_i) = \sigma^2 \text{ for all } i).$$

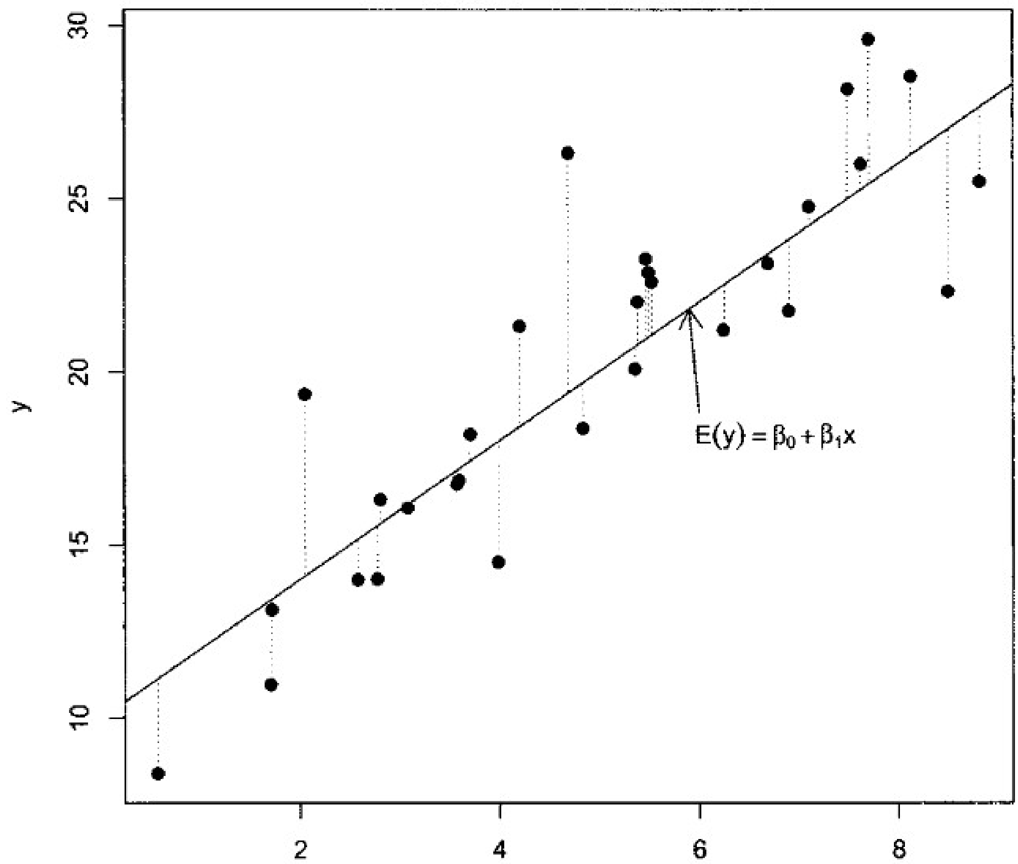
The errors are uncorrelated with each other.

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The regression model (1.1) is then

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

*



*

The normal equations [which determine the minimizer of (1.2)] can be shown (using multivariate calculus) to be

$$(X'X)\hat{\boldsymbol{\beta}} = X'\mathbf{y},$$

which implies that the least squares estimates satisfy

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}.$$

The fitted values are then

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} \equiv H\mathbf{y}, \tag{1.3}$$

where $H = X(X'X)^{-1}X'$ is the so-called "hat" matrix (since it takes $\mathbf{y}$ to $\hat{\mathbf{y}}$). The residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ thus satisfy

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X(X'X)^{-1}X'\mathbf{y} = (I - X(X'X)^{-1}X')\mathbf{y}, \tag{1.4}$$

or

$$\mathbf{e} = (I - H)\mathbf{y}.$$

*

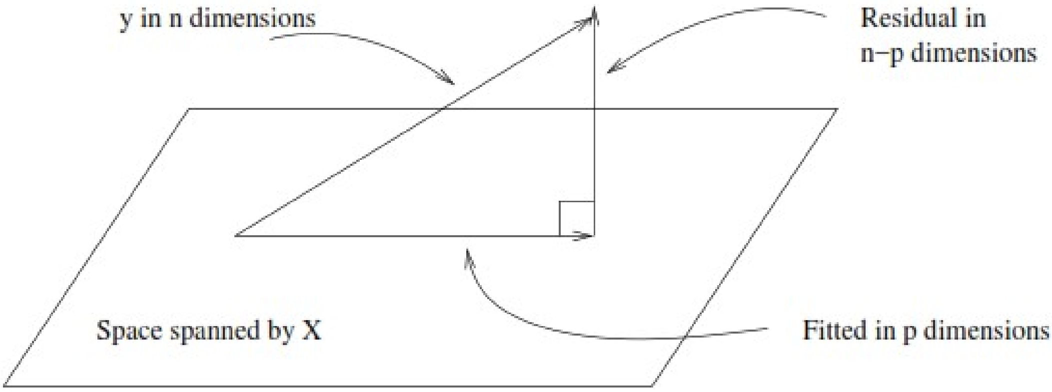


Figure 2.1: Geometric representation of the estimation β. The data vector Y is projected orthogonally onto the model space spanned by X. The fit is represented by projection $\hat{y} = X\hat{\beta}$ with the difference between the fit and the data represented by the residual vector $\hat{\varepsilon}$.