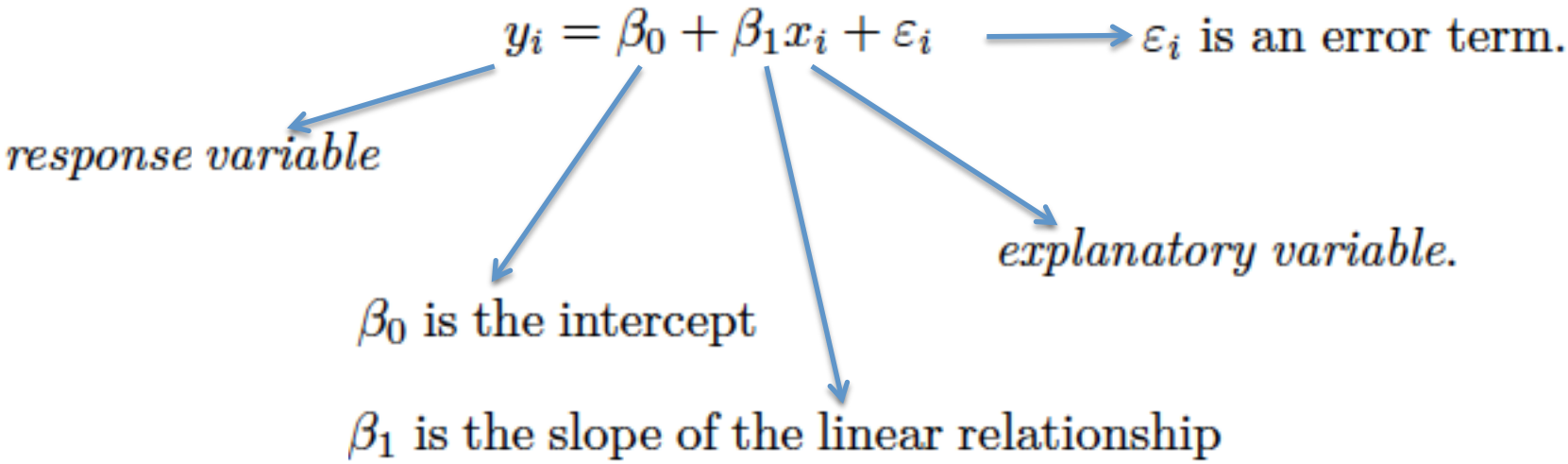
6 regresion lineal

Regresion lineal simple



El error tiene ddp normal de media cero y varianza σ^2

- Estimacion de minimos cuadrados

least squares estimation,

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ Minimizar la diferencia entre la prediccion y el valor observado

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\bar{y}$ and $\bar{x}$ are the means of the response and explanatory variable

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Estimacion de la varianza del error

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

Varianza del estimador del coef.

$$\mathrm{Var}(y_{\mathrm{pred}}) = \hat{\sigma}^2 \sqrt{\frac{1}{n} + 1 + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Varianza del predictor

Regresion lineal multiple

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \varepsilon_i.$$

$\varepsilon_i,\ i = 1, \ldots, n,$ i.i.d con ddp normal (0, σ^2)

$\mathsf{E}(y|x_1, \ldots, x_q) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$ Normal σ^2

$\beta_k,\ k = 1, \ldots, q,$ Coeficientes de regresion

β_0 Media general

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

$$\mathbf{y}^\top = (y_1, \ldots, y_n)$$
$$\beta^\top = (\beta_0, \beta_1, \ldots, \beta_q)$$
$$\varepsilon^\top = (\varepsilon_1, \ldots, \varepsilon_n)$$

design or *model matrix* $\mathbf{X}$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1q} \\ 1 & x_{21} & x_{22} & \ldots & x_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nq} \end{pmatrix}.$$

Variables categoricas se representan por un conjunto de vectores ortogonales

Coeficientes del intercept

- Estimador de minimos cuadrados

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \qquad \mathsf{E}(\hat{\beta}) = \beta$$

$$\mathsf{Var}(\hat{\beta}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

- Analisis de varianza

Table 6.3: Analysis of variance table for the multiple linear regression model.

Source of variation	Sum of squares	Degrees of freedom
Regression	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	q
Residual	$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	$n - q - 1$
Total	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{q1} \text{ and } \bar{y} = \sum_{i=1}^{n} y_i / n$$

- F-test $H_0 : \beta_1 = \cdots = \beta_q = 0.$

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / q}{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / (n - q - 1)}$$

- Estimacion de la varianza del ruido

$$\hat{\sigma}^2 = \frac{1}{n - q - 1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- Significacion $t_j = \hat{\beta}_j / \sqrt{\mathsf{Var}(\hat{\beta})_{jj}},$

- Diagnostico posthoc
  - Plot de los residuales contra cada variable explicativa
  - Plot de residuales contra los valores predichos
  - Plot de ajuste de los residuales a la distibribución normal

Caso 1

Freedman et al. (2001) give the relative velocity and the distance of 24 galaxies, according to measurements made using the Hubble Space Telescope – the data are contained in the **gamair** package accompanying Wood (2006), see Table 6.1. Velocities are assessed by measuring the Doppler red shift in the spectrum of light observed from the galaxies concerned, although some correction for 'local' velocity components is required. Distances are measured using the known relationship between the period of Cepheid variable stars and their luminosity. How can these data be used to estimate the age of the universe? Here we shall show how this can be done using simple linear regression.

```
install.packages(gamair)
data("hubble", package = "gamair")
View(hubble)
```

$$\text{velocity} = \beta_1 \text{distance} + \varepsilon.$$

This is essentially what astronomers call Hubble's Law and β_1 is known as Hubble's constant; β_1^{-1} gives an approximate age of the universe.

- Visualizar los datos

```
names(hubble)[3]<- "distance”
names(hubble)[2]<- "velocity"

plot(velocity ~ distance, data = hubble)
```

- Estimacion

```
> sum(hubble$distance * hubble$velocity) / sum(hubble$distance^2)
> hmod <- lm(velocity ~ distance - 1, data = hubble)  ## no intercept
>coef(hmod)
```

```
 > layout(matrix(1:2, ncol = 2))
> plot(velocity ~ distance, data = hubble)
> abline(hmod)
> plot(hmod, which = 1)
```

for the age of the universe. The Hubble constant itself has units of km $\times$ $\text{sec}^{-1} \times \text{Mpc}^{-1}$. A mega-parsec (Mpc) is 3.09×10^{19}km, so we need to divide the estimated value of β_1 by this amount in order to obtain Hubble's constant with units of sec^{-1}. The approximate age of the universe in seconds will then be the inverse of this calculation. Carrying out the necessary computations

```
> Mpc <- 3.09 * 10^19
> ysec <- 60^2 * 24 * 365.25
> Mpcyear <- Mpc / ysec
> 1 / (coef(hmod) / Mpcyear)
```

Caso 2

The data shown in Table 6.2 were collected in the summer of 1975 from an experiment to investigate the use of massive amounts of silver iodide (100 to 1000 grams per cloud) in cloud seeding to increase rainfall (Woodley et al., 1977). In the experiment, which was conducted in an area of Florida, 24 days were judged suitable for seeding on the basis that a measured suitability criterion, denoted *S-Ne*, was not less than 1.5. Here *S* is the 'seedability', the difference between the maximum height of a cloud if seeded and the same cloud if not seeded predicted by a suitable cloud model, and *Ne* is the number of hours between 1300 and 1600 G.M.T. with 10 centimetre echoes in the target; this quantity biases the decision for experimentation against naturally rainy days. Consequently, optimal days for seeding are those on which seedability is large and the natural rainfall early in the day is small.

On suitable days, a decision was taken at random as to whether to seed or not. For each day the following variables were measured:

View(clouds)

- Visualizacion

```
 > data("clouds", package = "HSAUR2")
> layout(matrix(1:2, nrow = 2))
> bxpseeding <- boxplot(rainfall ~ seeding, data = clouds,
+ ylab = "Rainfall", xlab = "Seeding")
> bxpecho <- boxplot(rainfall ~ echomotion, data = clouds,
+ ylab = "Rainfall", xlab = "Echo Motion")
```

```
 > layout(matrix(1:4, nrow = 2))
> plot(rainfall ~ time, data = clouds)
> plot(rainfall ~ cloudcover, data = clouds)
> plot(rainfall ~ sne, data = clouds, xlab="S-Ne criterion")
> plot(rainfall ~ prewetness, data = clouds)
```

- Outliers

```
> rownames(clouds)[clouds$rainfall %in% c(bxpseeding$out,
+ bxpecho$out)]
```

- Formula del modelo

```
R> clouds_formula <- rainfall ~ seeding +
+ seeding:(sne + cloudcover + prewetness + echomotion) +
+ time
```

- Matriz de diseño

```
Xstar <- model.matrix(clouds_formula, data = clouds)
```

- Estimacion

```
> clouds_lm <- lm(clouds_formula, data = clouds)
> class(clouds_lm)

>summary(clouds_lm)

   >betastar <- coef(clouds_lm)  #coeficientes de regresion

Vbetastar <- vcov(clouds_lm)

  sqrt(diag(Vbetastar))   # estándar errors
```

```
 > psymb <- as.numeric(clouds$seeding)
> plot(rainfall ~ sne, data = clouds, pch = psymb, xlab = "S-Ne criterion")
> abline(lm(rainfall ~ sne, data = clouds,
+ subset = seeding == "no"))
> abline(lm(rainfall ~ sne, data = clouds,
+ subset = seeding == "yes"), lty = 2)
> legend("topright", legend = c("No seeding", "Seeding"),
+ pch = 1:2, lty = 1:2, bty = "n")
```

- Calidad de la regresion

```
> clouds_resid <- residuals(clouds_lm)
> clouds_fitted <- fitted(clouds_lm)

> plot(clouds_fitted, clouds_resid, xlab = "Fitted values",
+ ylab = "Residuals", type = "n",
+ ylim = max(abs(clouds_resid)) * c(-1, 1))
> abline(h = 0, lty = 2)
> text(clouds_fitted, clouds_resid, labels = rownames(clouds))

> qqnorm(clouds_resid, ylab = "Residuals")
> qqline(clouds_resid)
```