Matrices and vectors

A matrix is a rectangular array of numbers. Here's an example:

$$\mathbf{A} = \begin{bmatrix} 2.3 & -1.4 & 17 \\ -22.5 & 0 & \sqrt{2} \end{bmatrix}.$$

This matrix has dimensions 2×3. The number of rows is first, then the number of columns.

We can write the $n \times p$ matrix $\mathbf{X}$ abstractly as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}.$$

Another notation that is common is $\mathbf{A} = [a_{ij}]_{n \times m}$ for an $n \times m$ matrix $\mathbf{A}$ with element a_{ij} in the i^{th} row and j^{th} column. The matrix $\mathbf{X}$ on the previous page would then be written $\mathbf{X} = [x_{ij}]_{n \times p}$.

If two matrices $\mathbf{A} = [a_{ij}]_{n \times m}$ and $\mathbf{B} = [b_{ij}]_{n \times m}$ have the same dimensions, you can add them together, element by element, to get a new matrix $\mathbf{C} = [c_{ij}]_{n \times m}$. That is, $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is the matrix with elements $c_{ij} = a_{ij} + b_{ij}$. For example,

$$\begin{bmatrix} -1 & -2 \\ 5 & 7 \\ -10 & 20 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} -1+1 & -2+2 \\ 5+3 & 7+4 \\ -10+1 & 20+2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 8 & 11 \\ -9 & 22 \end{bmatrix}.$$

Multiplying a matrix $\mathbf{A} = [a_{ij}]_{n\times m}$ by a number b yields the matrix $\mathbf{C} = \mathbf{A}b$ with elements $c_{ij} = a_{ij}b$. For example,

$$(-2)\begin{bmatrix} -1 & -2 \\ 5 & 7 \\ -10 & 20 \end{bmatrix} = \begin{bmatrix} -1(-2) & -2(-2) \\ 5(-2) & 7(-2) \\ -10(-2) & 20(-2) \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ -10 & -14 \\ 20 & -40 \end{bmatrix}.$$

The transpose of a matrix $\mathbf{A}'$ takes the matrix $\mathbf{A}$ and makes the rows the columns and the columns the rows. Precisely, if $\mathbf{A} = [a_{ij}]_{n\times m}$ then $\mathbf{A}'$ is the $m \times n$ matrix with elements $a'_{ij} = a_{ji}$. For example:

$$\text{If } \mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \text{ then } \mathbf{A}' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

Question: what is $(\mathbf{A}')'$?

A vector is a matrix with only one column or row, called a "column vector" or "row vector" respectively. Here's an example of each:

$$\mathbf{x} = \begin{bmatrix} 1 \\ -1 \\ 14 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 & -1 & 14 \end{bmatrix}.$$

Note that for these vectors, $\mathbf{x}' = \mathbf{y}$ and $\mathbf{y}' = \mathbf{x}$.

The product of an $1 \times n$ row vector and a $n \times 1$ column vector is the sum of the pairwise products of elements. So if $\mathbf{x} = [x_i]_{1\times n}$ and $\mathbf{y} = [y_i]_{n\times 1}$ then $\mathbf{xy} = \sum_{i=1}^{n} x_i y_i$.

For example, if $\mathbf{x} = [\ -1 \quad 2\ ]$ and $\mathbf{y} = \begin{bmatrix} 10 \\ -5 \end{bmatrix}$ then

$$\mathbf{x}\mathbf{y} = [\ -1 \quad 2\ ] \begin{bmatrix} 10 \\ -5 \end{bmatrix} = -1(10) + 2(-5) = -20.$$

The *inner product* of two $n \times 1$ column vectors $\mathbf{x}$ and $\mathbf{y}$ is the product

$$\mathbf{x}'\mathbf{y} = [\ x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_n\ ] \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Note that if $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is a point in the plane $\mathbb{R}^2$, then $\mathbf{x}'\mathbf{x} = x_1^2 + x_2^2$ is the square of the length of $\mathbf{x}$. That is, $||\mathbf{x}|| = \sqrt{\mathbf{x}'\mathbf{x}}$.

We are now ready to define general matrix multiplication. The product of an $n \times p$ matrix $\mathbf{A}$ and a $p \times m$ matrix $\mathbf{B}$ is the $n \times m$ matrix $\mathbf{C}$ with elements $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$. Let $\mathbf{A}$ be comprised of n $1 \times p$ row vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ and let $\mathbf{B}$ be comprised of m $p \times 1$ column vectors $\mathbf{b}_1, \ldots, \mathbf{b}_m$ like

$$\mathbf{A} = \begin{bmatrix} \cdots \mathbf{a}_1 \cdots \\ \cdots \mathbf{a}_2 \cdots \\ \vdots \\ \cdots \mathbf{a}_n \cdots \end{bmatrix}_{n \times p} \quad \text{and } \mathbf{B} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_m \\ \vdots & \vdots & & \vdots \end{bmatrix}_{p \times m} .$$

Then $c_{ij} = \mathbf{a}_i \mathbf{b}_j$:

$$\mathbf{C} = \begin{bmatrix} \mathbf{a}_1\mathbf{b}_1 & \mathbf{a}_1\mathbf{b}_2 & \cdots & \mathbf{a}_1\mathbf{b}_m \\ \mathbf{a}_2\mathbf{b}_1 & \mathbf{a}_2\mathbf{b}_2 & \cdots & \mathbf{a}_2\mathbf{b}_m \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_n\mathbf{b}_1 & \mathbf{a}_n\mathbf{b}_2 & \cdots & \mathbf{a}_n\mathbf{b}_m \end{bmatrix}.$$

For example, let $\mathbf{A} = \begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 2 & 0 & -2 \\ 0 & -5 & 7 \\ 1 & 0.5 & -4 \end{bmatrix}$.

Then

$$\begin{aligned} \mathbf{AB} &= \begin{bmatrix} 1(2) - 1(0) - 2(1) & 1(0) - 1(-5) - 2(0.5) & 1(-2) - 1(7) - 2(-4) \\ -3(2) - 1(0) + 5(1) & -3(0) - 1(-5) + 5(0.5) & -3(-2) - 1(7) + 5(-4) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 4 & 0 \\ -1 & 8.5 & -21 \end{bmatrix}. \end{aligned}$$

On the previous slide, does $\mathbf{BA}$ make sense? No. The rows of the first matrix must be the same length as the columns of the second. Note that, in general, $\mathbf{AB} \neq \mathbf{BA}$.

Define $\mathbf{I}_{n\times n}$ as

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Then

$$\mathbf{A}_{n\times p}\mathbf{I}_{p\times p} = \mathbf{A}_{n\times p} \text{ and } \mathbf{I}_{n\times n}\mathbf{A}_{n\times p} = \mathbf{A}_{n\times p},$$

for any $\mathbf{A}_{n\times p}$. The matrix $\mathbf{I}_{n\times n}$ is called the $n \times n$ *identity matrix*.

For example,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -2 \\ -3 & -1 & 5 \end{bmatrix}.$$

The *inverse* of a square $(p \times p)$ matrix $\mathbf{A}$ is the $p \times p$ matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_{p\times p}$. For example, if $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}$, then

$$\begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and so $\mathbf{A}^{-1} = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix}$. Note that we must have

$$\begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

as well.

There is an algorithm for finding the inverse of any size matrix but it is very computationally intensive, except for 2×2 matrices. Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Then

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

That is, switch the diagonal entries, multiply the off-diagonals by -1, and divide the works by $a_{11}a_{22} - a_{12}a_{21}$.

We can show that this is the inverse in class. Try it out on $\mathbf{A}$ on the previous slide.

Not every square matrix has an inverse. For example

$$\mathbf{A} = \begin{bmatrix} -1 & 2 \\ 2 & -4 \end{bmatrix},$$

does not. Try the formula on the previous slide out on this matrix. What happens?

Square matrices that do not have an inverse are said to be *singular*.

The two sample problem in terms of matrices

Recall the two-sample normal model with equal variances:

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{iid}{\sim} N(\mu_1, \sigma^2),$$

$$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{iid}{\sim} N(\mu_2, \sigma^2).$$

We can rewrite this as

$$Y_{ij} = \mu_i + e_{ij},$$

where

$$e_{ij} \overset{iid}{\sim} N(0, \sigma^2),$$

where $i = 1, 2$ indexes the group (1 or 2) and $j = 1, \ldots, n_i$ is the observation within the group.

Each piece of data Y_{ij} follows:

$$\begin{aligned}
Y_{11} &= \mu_1 + e_{11} \\
Y_{12} &= \mu_1 + e_{12} \\
Y_{13} &= \mu_1 + e_{13} \\
\vdots \quad & \vdots \quad \vdots \\
Y_{1n_1} &= \mu_1 + e_{1n_1} \\
Y_{21} &= \mu_2 + e_{21} \\
Y_{22} &= \mu_2 + e_{22} \\
Y_{23} &= \mu_2 + e_{23} \\
\vdots \quad & \vdots \quad \vdots \\
Y_{2n_2} &= \mu_2 + e_{2n_2}
\end{aligned}$$

Define the following vectors and matrices:

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_2} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \text{ and } \mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_2} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2n_2} \end{bmatrix}.$$

Then we can write the model succinctly as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{e}.$$

We'll show this on the board for $n_1 = n_2 = 3$.

It turns out the the MLE's for μ_1 and μ_2, namely $\hat{\mu}_1 = n_1^{-1} \sum_{j=1}^{n_1} y_{1j} = \bar{y}_{1\bullet}$ and $\hat{\mu}_2 = n_2^{-1} \sum_{j=1}^{n_2} y_{2j} = \bar{y}_{2\bullet}$ are obtained in matrix terms as

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

We'll show part of this:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}.$$

And so

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix}.$$

Also,

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_2} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix}.$$

Then

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix} = \begin{bmatrix} \bar{y}_{1\bullet} \\ \bar{y}_{2\bullet} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix},$$

as promised. The MLE of σ^2 in terms of matrices is

$$\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\mu}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\mu}})/(n_1 + n_2) = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\mu}}||^2/(n_1 + n_2).$$

What is the point? Although the two-sample normal model is fairly simple, very complex models with multiple predictors, both categorical and continuous, can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

including the simple linear regression model, multiple regression models, oneway and multiway ANOVA models, and ANCOVA models.

In

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

- $\mathbf{Y}$ is the $n \times 1$ *data vector*.
- $\mathbf{X}$ is the $n \times p$ *design matrix*. Often the i^{th} row of $\mathbf{X}$ is comprised of $p - 1$ measurements taken on the i^{th} subject in a study, e.g. the i^{th} row of an Excel spreadsheet, and an intercept term.
- $\boldsymbol{\beta}$ is the $p \times 1$ *coefficient vector*. For the two-sample model, $p = 2$ and $\boldsymbol{\beta} = (\mu_1, \mu_2)$.
- $\mathbf{e}$ is the $n \times 1$ error vector. All the elements of $\mathbf{e}$ are assumed to be *iid* $N(0, \sigma^2)$.

The j^{th} residual in group i is defined to be

$$r_{ij} = y_{ij} - \widehat{E(Y_{ij})} = y_{ij} - \hat{\mu}_i.$$

These can be plotted versus the group number $i = 1, 2$ to assess whether constant variance across groups is reasonable. A histogram of all $n = n_1 + n_2$ residuals can be used to assess the normality assumption. There are formal tests for both constant variance and for normality.

```
> skull <- c(english,celt)
> group <- c(rep("English",length(english)),rep("Celt",length(celt)))
> group <- factor(group)
> t.test(skull~group,var.equal=T)

        Two Sample t-test

data:  skull by group t = -7.6952, df = 32, p-value = 9.003e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -19.91906 -11.58094
```

```
sample estimates:
   mean in group Celt mean in group English
               130.75                  146.50

> fit1 <- lm(skull~group) #lm() is "linear model"
> summary(fit1)

Call: lm(formula = skull ~ group)

Residuals:
    Min      1Q  Median      3Q     Max
-14.500  -4.312   0.875   3.500  11.500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   130.750      1.489  87.798   <2e-16 ***
groupEnglish   15.750      2.047   7.695    9e-09 ***
---
F-statistic: 59.22 on 1 and 32 DF,  p-value: 9.003e-09

> r <- fit1$residuals # get residuals
> plot(group,r) # gives boxplots instead of scatterplot
> hist(r) # looks roughly normal
```

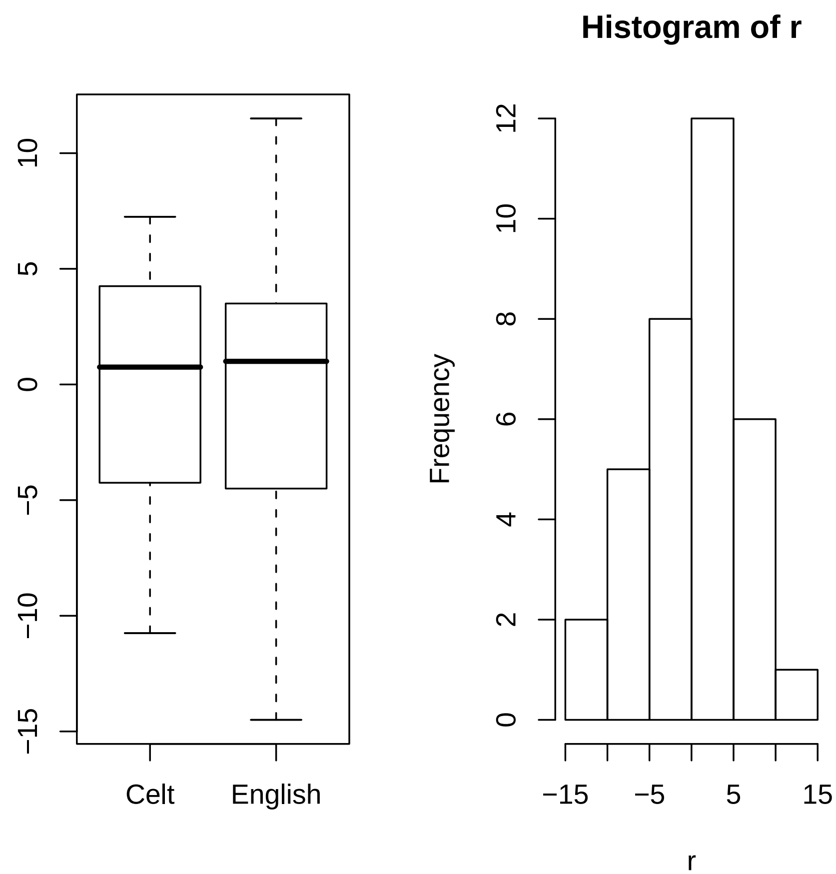

Figure 1: Boxplots show roughly the same spread; histogram is roughly ‘normal’ looking.

The `lm()` fit to the data fits a different, but equivalent model:

$$\text{Celts} : Y_{1,1}, \ldots, Y_{1,16} \stackrel{iid}{\sim} N(\mu, \sigma^2),$$

$$\text{English} : Y_{2,1}, \ldots, Y_{2,18} \stackrel{iid}{\sim} N(\mu + \delta, \sigma^2),$$

reports estimates of the mean Celtic headbreadth μ, and the mean difference for Englishmen δ. Of interest in this reparameterized model is $H_0 : \delta = 0$, which is the same as testing $H_0 : \mu_1 = \mu_2$ in the two-sample model.

The **simple linear regression model** stipulates

$$Y_i = \beta_0 + \beta_1 x_i + e_i,$$

where $i = 1, \ldots, n$. Here we can write

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \text{ and } \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

The model is succinctly written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

As in the two-sample model, the MLE's $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ are given by

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2.$$

For the i^{th} observation, the *fitted value* $\hat{y}_i$ is the point on the line at x_i, namely $\hat{\beta}_0 + \hat{\beta}_1 x_i$. So at value x_i we *see* the observed value y_i, but estimate the underlying mean $\hat{y}_i = \widehat{E(Y_i)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Since the MLE's $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased,

$$E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_i = \beta_0 + \beta_1 x_i.$$

That is, $E(\hat{y}_i) = \beta_0 + \beta_1 x_i$.

The i^{th} *residual* r_i estimates the unknown $e_i = y_i - [\beta_0 + \beta_1 x_i]$ by $r_i = y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i] = y_i - \widehat{E(Y_i)}$.

These are used to check model assumptions: (1) that the $e_1, \ldots, e_n \overset{iid}{\sim} N(0, \sigma^2)$, and (2) that the mean changes linearly with x: $E(Y_i) = \beta_0 + \beta_1 x_i$.

A histogram of $r_1, \ldots, r_n$ should be roughly symmetric and unimodal, i.e. "normal looking." A plot of r_i versus x_i should show no discernable pattern and roughly constant spread.

```
> fit2 <- lm(score~ses)
> plot(fit2) # get a variety of diagnostic plots automatically
> r <- fit2$residuals
> plot(ses,r)
> hist(r)
```

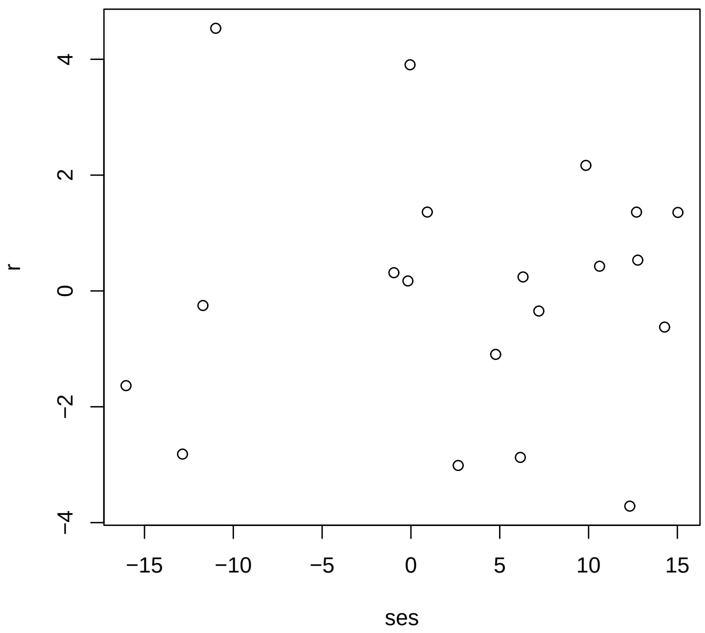

Figure 2: Residual plot shows no discernable pattern.

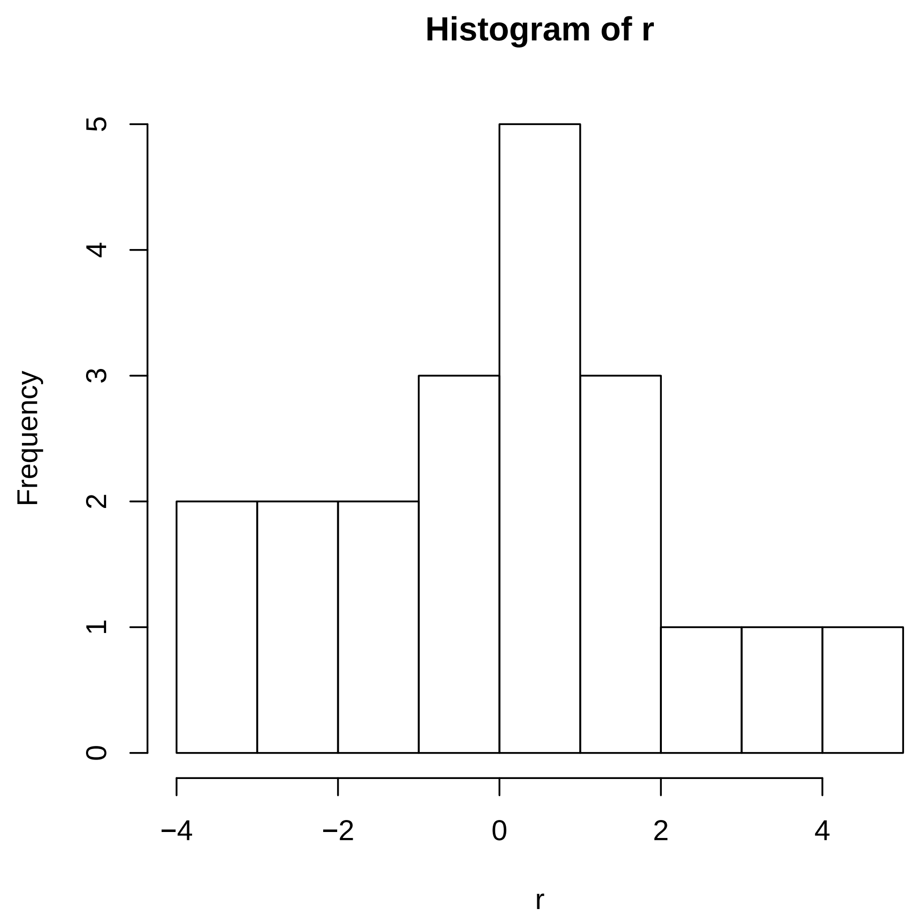

Figure 3: Histogram of residuals is roughly "normal looking."