STAT 210
Applied Statistics and Data Analysis
Linear Regression I: Introduction

Joaquin Ortega

Fall 2020

Linear Regression

Linear Regression

We now consider the statistical problem of building models that relate several variables. We start with simple models that link two variables to introduce the basic notions and set up the notation, and then we proceed to more complex models.

The simplest model concerns a pair of continuous variables X and Y.

Suppose we have a joint sample

$$(X_1, Y_1), (X_2, Y_2) \ldots, (X_n, Y_n)$$

and we want to determine whether there exists a relationship between them.

Simple Linear Regression

The simplest relation is a linear model such as

$$Y = \beta_0 + \beta_1 X \tag{1}$$

In this model, Y is the **response** or dependent variable and X is a (continuous) **explanatory** or independent variable, also known as a **regressor**.

There are two **parameters** in the model, the slope β_1 and the intercept β_0.

Regression analysis is used for building models like this or with a more complex structure.

The name **regression** is due to Sir Francis Galton, who showed that descendants of tall parents tend to be smaller than their parents, while those of short parents tend to be taller. He called this *regression towards mediocrity*.

Simple Linear Regression

Equation (1) is deterministic: if we know the value of X and the parameters of the model, we know the exact value of Y.

The models that will concern us will have a random component:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where ϵ is known as the error term.

This means that there is not an exact linear relationship between the variables X and Y.

As is usual, we will denote random variables and functions by capital letters, $Y_i, X_i, i = 1, \ldots, n$, and their values will be represented by small case letters $y_i, x_i, i = 1, \ldots, n$.

Simple Linear Regression

Since we have a sample of values from both variables $(X_i, Y_i), i = 1, \ldots, n$, the model is usually written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

This model is

- **simple**, because it has only one regressor or independent variable,
- **linear**, because it is linear in the parameters: none of the parameters appears as an exponent or raised to a power or multiplied by another parameter,

and is also linear on the variables because the predictor variable only appears raised to the power 1.

Simple Linear Regression

In Statistics, linear models are models that are *linear in the parameters.* For example,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$
$$Y_i = \beta_0 + \beta_1 \log(X_i) + \epsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_i^2 + \beta_2 e^{X_i} + \epsilon_i$$

are all linear models while

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \epsilon_i$$
$$Y_i = \frac{\beta_0}{1 + e^{\beta_1 X_i}} + \epsilon_i$$

are not.

Simple Linear Regression

Going back to model (2), ϵ_i is a random variable while X_i is not.

Y_i is also a random variable since it is a function of ϵ_i, but it also depends on X_i.

We usually assume that the ϵ_i are centered, $E[\epsilon_i] = 0$ and have equal variance $Var(\epsilon_i) = \sigma^2, i = 1, \ldots, n$. We will also assume that they follow a Gaussian distribution and are independent.

The expected value of Y given X is

$$\begin{aligned} E[Y|X] &= E[\beta_0 + \beta_1 X + \epsilon_i] \\ &= \beta_0 + \beta_1 E[X] + E[\epsilon_i] \\ &= \beta_0 + \beta_1 X. \end{aligned}$$

Simple Linear Regression

The distribution of Y **when X is known** is Gaussian with mean $\beta_0 + \beta_1 X$ and variance σ^2.

The slope β_1 represents the expected change in Y when X changes one unit.

When $\beta_1 = 0$, the response Y is independent of the explanatory variable X.

When $\beta_1 > 0$, X and Y have a positive linear relation, so Y increases with X, while if $\beta_1 < 0$, this linear relation is negative: when X increases, Y decreases.

Simple Linear Regression

The problem we want to solve is the estimation of the parameter for the models from a sample of values $(x_1, y_1), \ldots, (x_n, y_n)$.

We can write the relation as a system of linear equations

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \vdots \quad &\quad \vdots \quad \quad \vdots \quad \quad \vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n. \end{aligned}$$

Simple Linear Regression

In matrix notation this can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

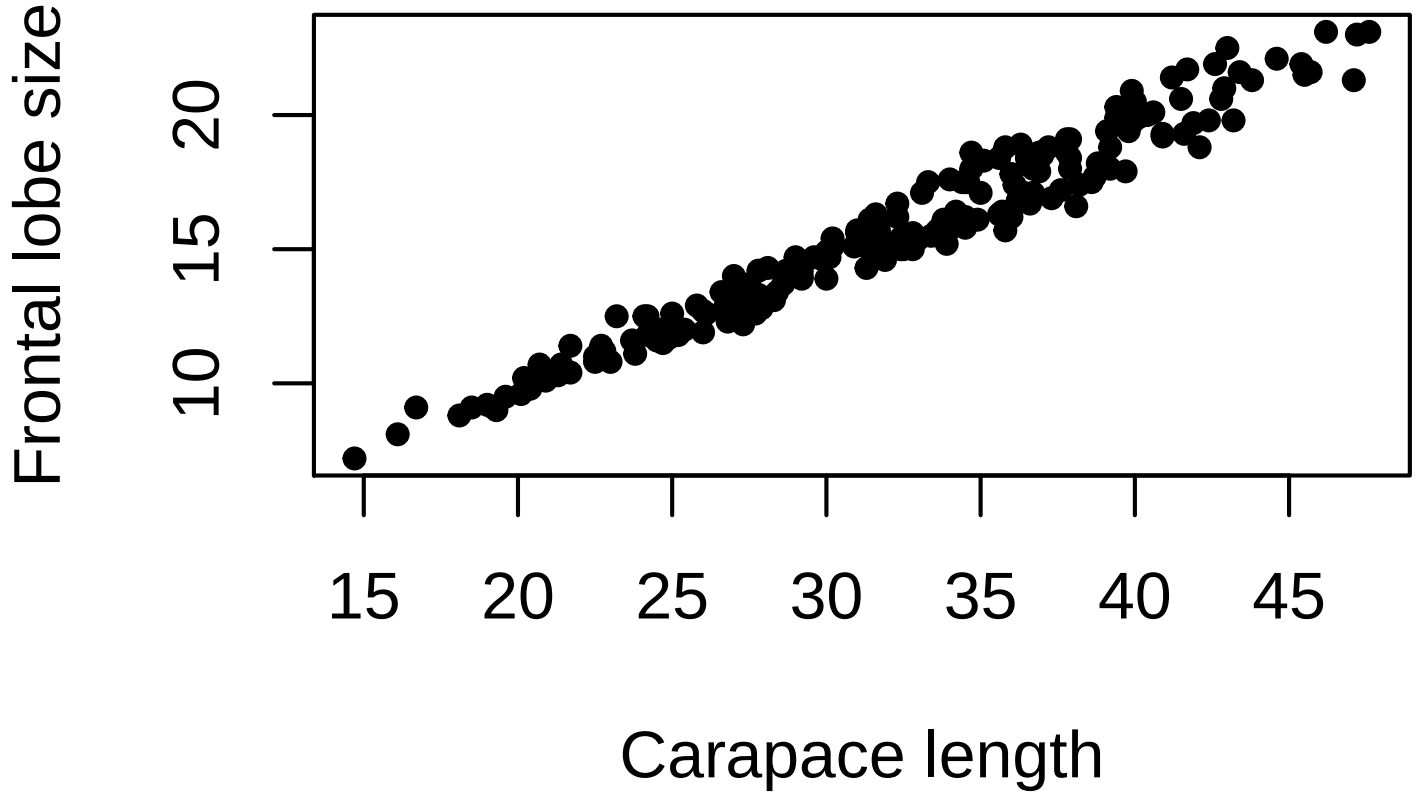
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}; \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

$\mathbf{X}$ is known as the **design matrix**, while $\boldsymbol{\beta}$ is the vector of parameters.

Example 1

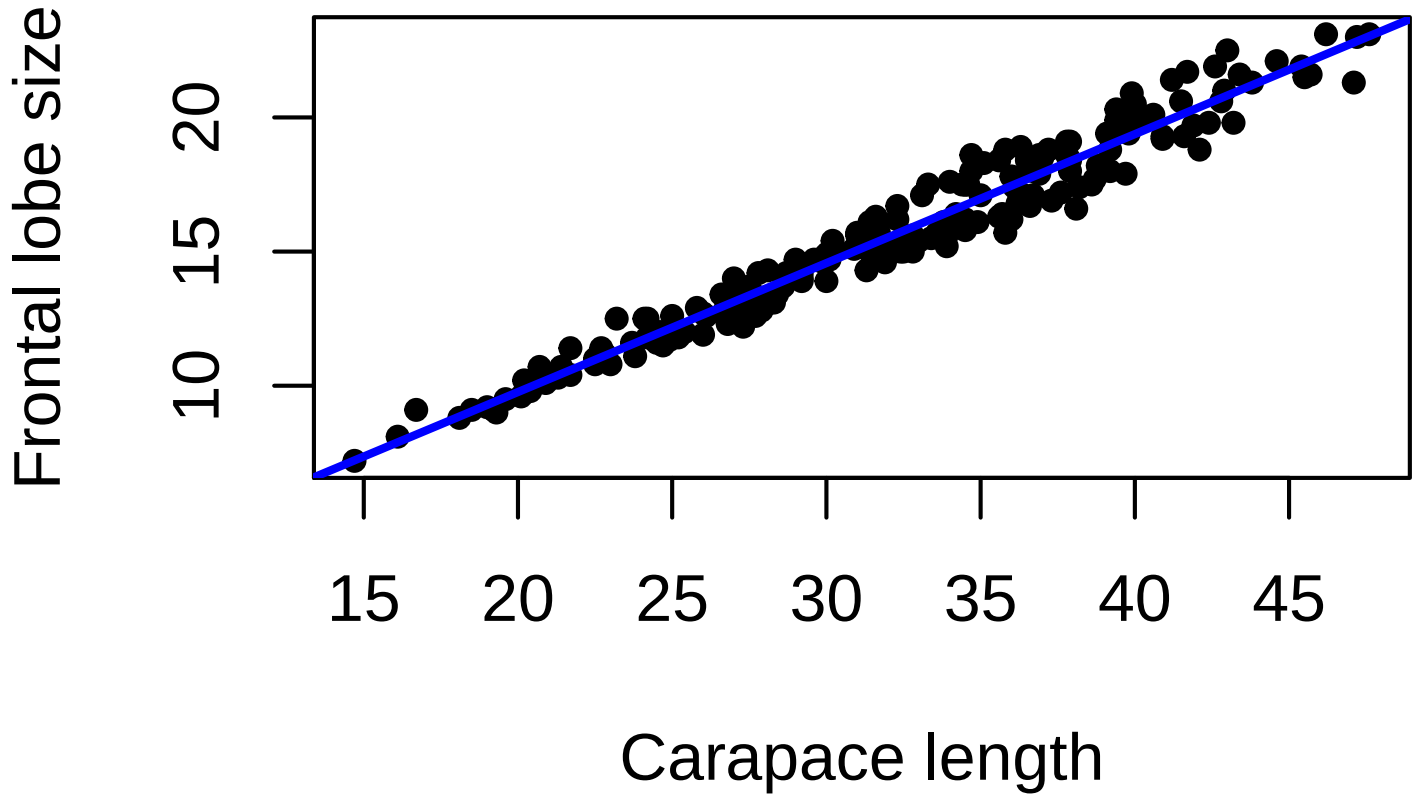
The data set `crabs` in the `MASS` package has data on five morphological variables of two color forms of crabs, 50 individuals for each color form and sex.

```
library(MASS); attach(crabs)
plot(CL,FL, pch=20, xlab='Carapace length', ylab='Frontal lobe size')
```



Example 1

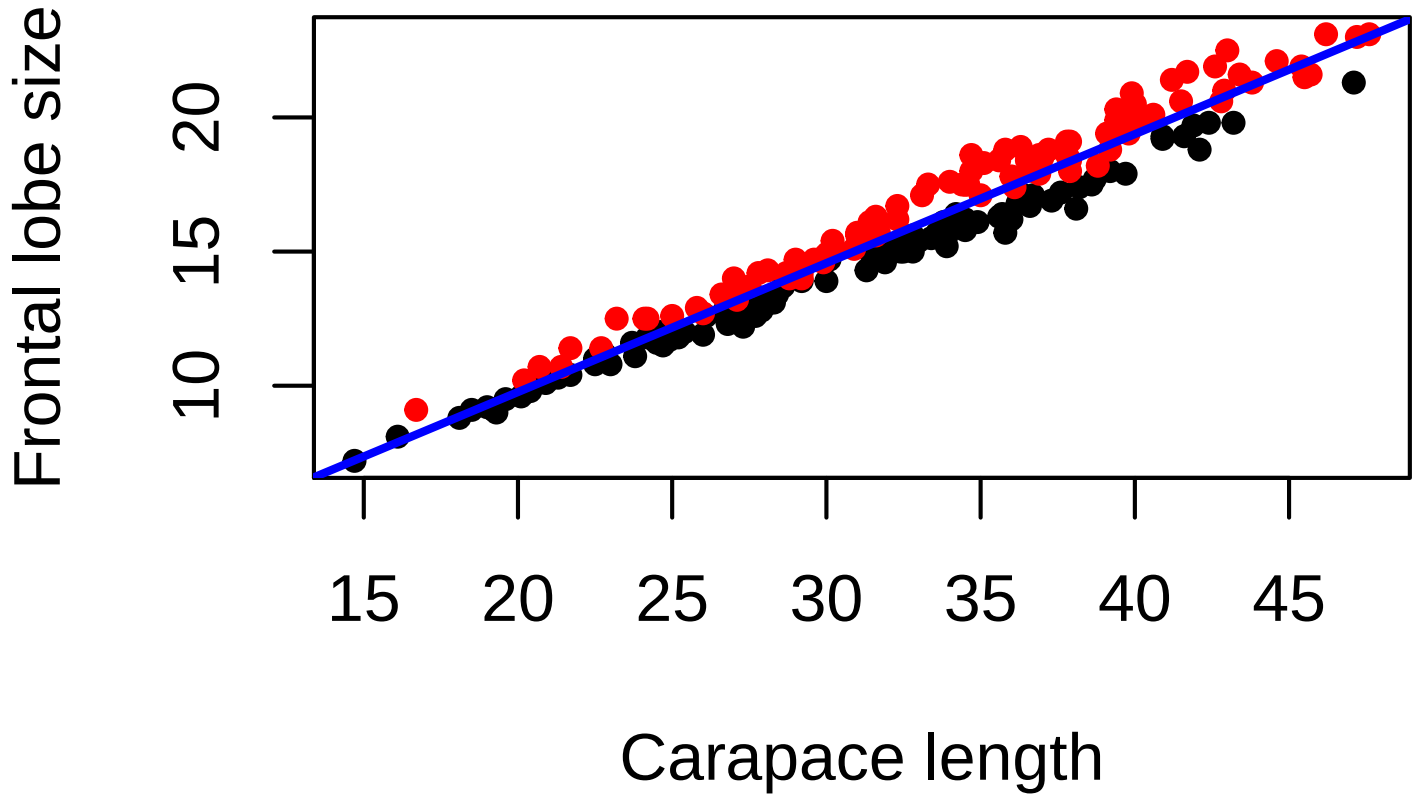
```
lm1 <- lm(FL~CL)
abline(lm1, lw=2, col='blue')
```



The line seems to describe the relationship between these two variables well, but if we add some color to the species

Example 1

```
lm1 <- lm(FL~CL)
abline(lm1, lw=2, col='blue')
```



We see that perhaps a better approach would be to separate the two species and fit different models for each.

Estimation

Estimation

We seek the straight line with the 'best fit' to the data. In our case, 'best' means minimizing the errors in some sense.

The model we want to fit is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where (x_i, y_i) are the observed values.

The errors are the differences between the observed values and the values that the model predicts: $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$.

Since these errors will be of different signs, we adopt the least-squares criterion for choosing the parameter values. If we denote by $\hat{\beta}_0, \hat{\beta}_1$ the estimated parameters, we want

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3}$$

Estimation

The sum on the right of (3) is known as the **error sum of squares** SSE:

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

and in matrix notation

$$SSE = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Once the parameters are estimated, we have the regression line

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

which can be used to predict values.

Estimation

The fitted values are given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and the residuals are given by

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Residual analysis is a fundamental step in model fitting to verify that assumptions are satisfied and also that as much variability as possible has been accounted for in the model.

Estimation

Derivation of the estimated values for the βs

Going back to equation (3) we want to minimize SSE. For this, we take partial derivatives wrt the parameters and set them to zero:

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{\partial SSE}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \qquad (4)$$

These are known as the **normal equations**. From the first equation we get that

$$\beta_0 = \frac{1}{n}\sum_i y_i - \beta_1 \frac{1}{n}\sum_i x_i = \bar{y} - \beta_1 \bar{x}.$$

Estimation

Equation (4) gives

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0. \tag{5}$$

Replacing $\beta_0 = \bar{y} - \beta_1 \bar{x}$ in this equation we get

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Rearranging terms

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \beta_1 \Big(\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \Big)$$

whence

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}.$$

Estimation

Therefore, the least squares estimators for the model parameters are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}. \tag{7}$$

Back to Example 1

Let's go back to example 1 and see what information we can get from the `lm` function:

```
summary(lm1)
```

```
##
## Call:
## lm(formula = FL ~ CL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86395 -0.51746 -0.02826  0.50456  1.77009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15316    0.23477   0.652    0.515
## CL           0.48060    0.00714  67.313   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.717 on 198 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9579
## F-statistic:  4531 on 1 and 198 DF,  p-value: < 2.2e-16
```

Example 1

The summary starts recalling the `call` for the `lm` function. Then we get an overview of the values for the residuals: Minimum, maximum, first and third quartiles, and median. These values give an idea of the symmetry of their distribution, which, in this case, looks good.

Next, we have a `Coefficients` matrix, with the values for the slope (0.15316) and the intercept (0.48060) in the `Estimate` column, and their standard errors in the `Std. Error` column. We will consider the meaning of the other two columns later on.

Example 1

The `lm` function produces an object of class `lm`, which is a list with 12 components

```
names(lm1)
```

```
##  [1] "coefficients"  "residuals"     "effects"
##  [4] "rank"          "fitted.values" "assign"
##  [7] "qr"            "df.residual"   "xlevels"
## [10] "call"          "terms"         "model"
```

Example 1

A more detailed look:

```
str(lm1)
```

```
## List of 12
##  $ coefficients : Named num [1:2] 0.153 0.481
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "CL"
##  $ residuals    : Named num [1:200] 0.2092 -0.052 -0.0845 -0.2132 -0.1093 ..
##   ..- attr(*, "names")= chr [1:200] "1" "2" "3" "4" ...
##  $ effects      : Named num [1:200] -220.3769 48.2644 -0.0913 -0.2206 -0.116
##   ..- attr(*, "names")= chr [1:200] "(Intercept)" "CL" "" "" ...
##  $ rank         : int 2
##  $ fitted.values: Named num [1:200] 7.89 8.85 9.28 9.81 9.91 ...
##   ..- attr(*, "names")= chr [1:200] "1" "2" "3" "4" ...
##  $ assign       : int [1:2] 0 1
##  $ qr           :List of 5
##   ..$ qr   : num [1:200, 1:2] -14.1421 0.0707 0.0707 0.0707 0.0707 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:200] "1" "2" "3" "4" ...
##   .. .. ..$ : chr [1:2] "(Intercept)" "CL"
##   .. ..- attr(*, "assign")= int [1:2] 0 1
##   ..$ qraux: num [1:2] 1.07 1.13
##   ..$ pivot: int [1:2] 1 2
##   ..$ tol  : num 1e-07
##   ..$ rank : int 2
##   ..- attr(*, "class")= chr "qr"
##  $ df.residual  : int 198
```

Example 1

We can retrieve the coefficients for the model with

```
lm1$coefficients
```

```
## (Intercept)          CL
##   0.1531552   0.4805982
```

or also

```
coef(lm1)
```

```
## (Intercept)          CL
##   0.1531552   0.4805982
```

Example 1

It is also useful to look at the structure of `summary(lm1)`

```
str(summary(lm1))
```

```
## List of 11
##  $ call         : language lm(formula = FL ~ CL)
##  $ terms        :Classes 'terms', 'formula'  language FL ~ CL
##   .. ..- attr(*, "variables")= language list(FL, CL)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:2] "FL" "CL"
##   .. .. .. ..$ : chr "CL"
##   .. ..- attr(*, "term.labels")= chr "CL"
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(FL, CL)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. ..- attr(*, "names")= chr [1:2] "FL" "CL"
##  $ residuals    : Named num [1:200] 0.2092 -0.052 -0.0845 -0.2132 -0.1093 ...
##   ..- attr(*, "names")= chr [1:200] "1" "2" "3" "4" ...
##  $ coefficients : num [1:2, 1:4] 0.15316 0.4806 0.23476 0.00714 0.65238 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "(Intercept)" "CL"
##   .. ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
##  $ aliased      : Named logi [1:2] FALSE FALSE
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "CL"
##  $ sigma        : num 0.717
##  $ df           : int [1:3] 2 198 2
##  $ r.squared    : num 0.958
##  $ adj.r.squared: num 0.958
##  $ fstatistic   : Named num [1:3] 4531 1 198
##   ..- attr(*, "names")= chr [1:3] "value" "numdf" "dendf"
##  $ cov.unscaled : num [1:2, 1:2] 1.07e-01 -3.18e-03 -3.18e-03 9.92e-05
```

Residuals

Residuals

The difference between the observed values y_i, and the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, n$ are the residuals:

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

The model we have fitted is the one that minimizes the sum of squares of these residuals. Observe that

$$\begin{aligned} \sum_{i=1}^{n} \hat{\varepsilon}_i &= \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\ &= n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 n\bar{x} = 0. \end{aligned} \tag{8}$$

Another formula for the parameters

Another formula for the parameters

Recall that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_1^2 - n(\bar{x})^2}. \tag{7}$$

There is another expression for this parameter. Observe that

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i - \bar{y}\sum_{i=1}^n x_i + n\bar{x}\bar{y} \\
&= \sum_{i=1}^n x_i y_i - \bar{x} n\bar{y} - \bar{y} n\bar{x} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y},
\end{aligned}$$

so the numerator in (7) can be replaced by $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$.

Another formula for the parameters

On the other hand,

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

Another formula for the parameters

Using these two relations in (7) we get that

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\end{aligned} \tag{9}$$

We see that the numerator is an estimator of the covariance between x_i and y_i, $i = 1, \ldots, n$ while the denominator is an estimator of the 'variance' of the $x_i, i = 1, \ldots n$.

Properties of the Regression Line

Properties of the Regression Line

We have already seen in (8) that $\sum_i \hat{\epsilon}_i = 0$. Other properties are:

1. $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$.
2. $\sum_{i=1}^n x_i \hat{\epsilon}_i = 0$.
3. $\sum_{i=1}^n \hat{y}_i \hat{\epsilon}_i = 0$.
4. The regression line always goes through $(\bar{x}, \bar{y})$.

Proofs The first property follows from (8):

$$0 = \sum_{i=1}^n \hat{\epsilon}_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{y}_i.$$

Properties of the Regression Line

For the second property, from (5) we have

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) \\ &= \sum_{i=1}^{n} x_i \hat{\epsilon}_i. \end{aligned}$$

Next,

$$\begin{aligned} \sum_{i=1}^{n} \hat{y}_i \hat{\epsilon}_i &= \sum_{i=1}^{n} \hat{\epsilon}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \hat{\beta}_0 \sum_{i=1}^{n} \hat{\epsilon}_i + \hat{\beta}_1 \sum_{i=1}^{n} \hat{\epsilon}_i x_i = 0 \end{aligned}$$

by (8) and property 2.

Properties of the Regression Line

Finally, the regression line is given by $y = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, therefore

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

and if $x = \bar{x}$, we see that $y = \bar{y}$.

Observe that properties 2 and 3 imply that, as vectors,

$$\mathbf{x} \cdot \hat{\epsilon} = 0, \qquad \hat{\mathbf{y}} \cdot \hat{\epsilon} = 0$$