Linear regression models with linear algebra and R

Juho Kopra

2023-02-13

1. Scatter plot and one linear predictor

Let's consider following data set, and draw a scatter plot out of it.

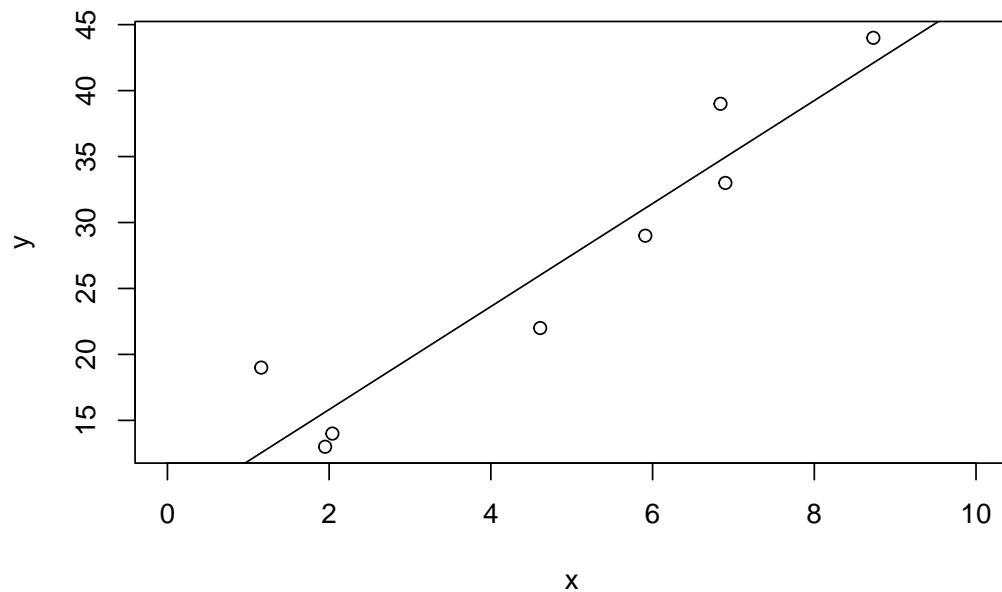| x	y
6.84	39
8.73	44
6.90	33
1.16	19
1.95	13
4.61	22
2.04	14
5.91	29

The scatterplot becomes

Figure 1: Scatterplot and a linear model with one predictor.

The coefficients of a linear regression fitted to that data set are $\beta_0 = 8.01$ and $\beta_1 = 3.90$.

Definition

Importantly, a linear regression model with one variable is a statistical model which can be formulated as

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

and where $i = 1, \ldots, n$ and $\varepsilon_i \sim N(0, \sigma^2)$.

There the following terminology takes place

- Y_i is dependent
- β_0 is an intercept
- β_1 is a slope
- β_0 and β_1 are regression coefficients or model parameters
- ε_i is an error term
- x_i is the value of predictor aka independent aka covariate aka regressor

In equation (1) we did not write estimated values of parameters down to the equation but used β-coefficients instead. When we write it down with β-coefficients we call it general form of linear model.

Linear model with multiple predictors $X_1, X_2, \ldots, X_k$ it can be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik} + \varepsilon_i \tag{2}$$

and where $i = 1, \ldots, n$ and $\varepsilon_i \sim N(0, \sigma^2)$.

The fitting of linear model

We want to minimize the sum of squares of error terms ε_i^2, which can be written as $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$. Note that error terms are in the same direction with Y-axis (and thus the same direction with Y_i).

We will not go into technical details of minimization here.

Using R we can fit the above model as follows. Assume that we have already read in data into object `dat`. We use `lm` function to fit the **l**inear **m**odel. The resulting calculation we store into object named `m1`. The data object `dat` has columns named `x` and `y` but the names could be something else as well. The first argument of `lm` is formula, which is special type in R. Formula is handy for fitting many types of models. On the left hand side of the formula one writes the dependent (e.g. what is Y_i), then tilde (~) and then predictor variable(s). For multiple predictors one need to put a plus sign (+) between the predictors.

```
m1 <- lm(y ~ x, data=dat)
m1
```

```
##
## Call:
## lm(formula = y ~ x, data = dat)
##
## Coefficients:
## (Intercept)            x
##        8.01         3.90
```

2. Properties of linear models and interpretation

Properties of linear models

To have an interpretation for β_0 and β_1 we note that:

Expected value of Y_i is if we consider x_i being known and fixed:

$$
\begin{aligned}
E(Y_i|x_i) &= E(\beta_0 + \beta_1 x_i + \varepsilon_i) \\
&= E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i) \\
&= E(\beta_0) + x_i E(\beta_1) + E(\varepsilon_i) \\
&= \beta_0 + \beta_1 x_i
\end{aligned}
$$

because x_i is constant and $E(\varepsilon_i) = 0$.

Shortly put, the above calculation gives $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$, which is called a systematic part of linear model. If we replace the β-notation with their estimates, then we can use that equation to calculate fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ Fitted values are calculated for the same x_i values which were in the original data.

```
fitted(m1)
```

```
##      1      2      3      4      5      6      7      8
## 34.716 42.095 34.951 12.541 15.625 26.010 15.976 31.085
```

Prediction can be obtained by the same equation but using any value. Here we predict new values using x=10 and x=11.

```
predict(m1,newdata = data.frame(x=c(10,11)))
```

```
##      1      2
## 47.053 50.958
```

Interpretation

We want to interpret what happens in the population of our study. That is why we use equation of expected value of Y_i, that is $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$ to give interpretation.

An interpretation of β_1 becomes available as

$$
\begin{aligned}
E(Y_i|x_i) - E(Y_i|x_i + 1) &= \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1(x_i + 1)) \\
&= \beta_1 x_i - \beta_1(x_i + 1) \\
&= \beta_1 x_i - \beta_1 x_i + \beta_1 \cdot 1 \\
&= \beta_1
\end{aligned}
$$

Thus, if e.g. $\beta_1 = 3.90415$ then the expected value of Y_i increases by 3.90415 units of y if x_i becomes increased with 1 unit of x.

An interpretation of intercept term β_0 becomes available if we set $x_i = 0$ since that removes β_1 from the equation. Thus $E(Y_i|x_i = 0) = \beta_0 + \beta_1 \cdot 0 = \beta_0$.

3. Categorical predictors

The predictors may as well be categorical ones. For example, we can model the mice weights where mice have two groups in a study. We are used to do that with a t-test but here we formulate it using linear regression. The results are the very same as with t-test!
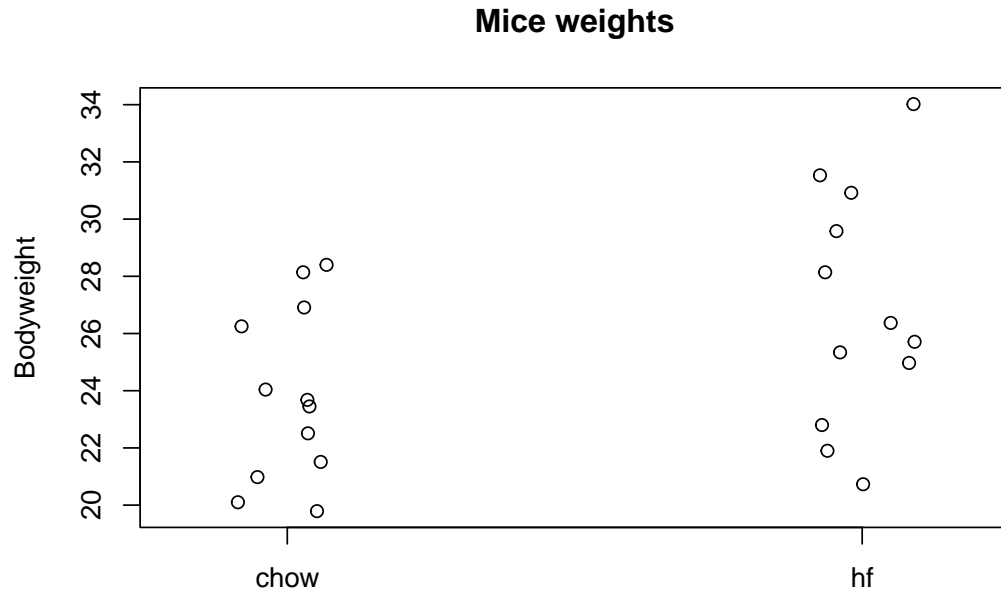
Mice weights

Figure 2: Mice weight with jitter. Example borrowed from Mike Love.

```r
m1 <- lm(Bodyweight ~ Diet, data=dat)
```

```r
summary(m1)
```

```
##
## Call:
## lm(formula = Bodyweight ~ Diet, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.104 -2.436 -0.414  2.834  7.186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23.81       1.04   22.91   <2e-16 ***
## Diethf          3.02       1.47    2.06    0.052 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.6 on 22 degrees of freedom
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.123
## F-statistic: 4.22 on 1 and 22 DF,  p-value: 0.0519
```

Above, the `(Intercept)` represents the average of group `chow` that is 23.81333 and the `Diethf` is parameter comparing the differences of means between the two groups. Thus, p-value $p = 0.052$ tells us that these two groups do not differ from each other.

```r
t.test(Bodyweight ~ Diet, data=dat, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  Bodyweight by Diet
## t = -2.06, df = 22, p-value = 0.052
## alternative hypothesis: true difference in means between group chow and group hf is not equal to 0
## 95 percent confidence interval:
##  -6.069152  0.027485
## sample estimates:
## mean in group chow    mean in group hf
##             23.813              26.834
```

Dummy variables and indicator function Let's familiariase ourselves with dummy variables. Dummy variable is a variable which can have either value 0 or 1. For the example above, the dummy variable can be used to indicate that diet of a mouse is `hf` instead of `chow`. Thus, let's create a dummy variable so that it takes a value 1 if diet is `hf` and value 0 if diet is `chow`.

In mathematical formulations, a useful way of reminding ourselves that a predictor is a dummy variable is to use an indicator function. Indicator function $\mathbf{1}(x)$ takes value 1 if the condition x is true, and value 0 otherwise. The notation of indicator variable nicely reminds us that it will give value 1 or 0.

Indicator function is

$$\mathbf{1}(x) = \begin{cases} 1 & \text{if condition } x \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

For example we may write

$$Y_i = \beta_0 + \beta_1 \mathbf{1}(X_i = \texttt{hf}) + \varepsilon_i$$

where Y_i and x_i stand for values of `Bodyweight` and `Diet` respectively, and $i = 1, \ldots, n$ and $\varepsilon_i \sim N(0, \sigma^2)$. In the above, the $\mathbf{1}(X_i = \texttt{hf})$ gives a value 1 if `Diet` is `hf`. Otherwise it gives 0.

Thus, the expected value of Y_i is β_0 if diet is `chow`. If the diet is `hf`, the expected values of Y_i is $\beta_0 + \beta_1$.

4. Basics of linear algebra

Linear algebra is a branch of mathematics where instead of numbers (aka scalars) we operate on vectors and matrices. Linear algebra is a useful tool for solving systems of linear equations, and thus suitable for handling mathematics behind linear regression. Next, we go through some of the key concepts so that we can move forward to understand how linear regression can be solved.

Transpose of matrix $\mathbf{A}$ is denoted as $\mathbf{A}^T$

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \implies \mathbf{A}^T = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}$$

Above matrix A is n times p matrix, where $n = 2, p = 3$.

```
A <- matrix(c(1,2,
              3,4,
              5,6),nrow=3,ncol=2,byrow=T)
A
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
## [3,]    5    6
```

```
t(A)
```

```
##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
```

A matrix with only one column is called a vector.

$$d = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$$

A transpose of a vector is a row vector.

$$d^T = \begin{bmatrix} 1 & 3 & 5 \end{bmatrix}$$

A cross product (aka matrix multiplication) of row vector with a column vector is

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} = \begin{bmatrix} 1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 \end{bmatrix} = \begin{bmatrix} 32 \end{bmatrix}$$

```
d <- c(4,5,6)
d
```

```
## [1] 4 5 6
```

```
t(d)
```

```
##      [,1] [,2] [,3]
## [1,]    4    5    6
```

```
t(t(d))
```

```
##      [,1]
## [1,]    4
## [2,]    5
## [3,]    6
```

```
c(1,2,3) %*% d
```

```
##      [,1]
## [1,]   32
```

Identity matrix is **I** has 1 on diagonal and 0 elsewhere. E.g.

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

```
diag(3)
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

If matrix has the same number of rows and columns, it is called a square matrix.

If a square matrix **B** has determinant, which is not zero, then **B** is invertible, which means that it is possible to calculate an inverse of matrix **B** that is $\mathbf{B}^{-1}$.

```
# Let's create a square matrix
B <- t(A) %*% A
B
```

```
##      [,1] [,2]
## [1,]   35   44
## [2,]   44   56
```

```
det(B) #determinant is not zero => invertible
```

```
## [1] 24
```

```
solve(B)
```

```
##          [,1]    [,2]
## [1,]  2.3333 -1.8333
## [2,] -1.8333  1.4583
```

For invertible matrix **B** it follows that $\mathbf{BB}^{-1} = \mathbf{I}$ and also $\mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$.

Solving systems of equations using linear algebra

Matrices can be used to represent a systems of equations. To understand the basic concepts, let's have a look at following system of equations:

$$a + b + c = 6$$
$$3a - 2b + c = 2$$
$$2a + b - c = 1$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 3 & -2 & 1 \\ 2 & 1 & -1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \\ 1 \end{pmatrix}$$

Taking the inverse of square matrix and multiplying the equation from the left hand side with that solves the system.

$$\implies \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 3 & -2 & 1 \\ 2 & 1 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 6 \\ 2 \\ 1 \end{pmatrix}$$

The same as above with R

```
A <- matrix(c(1, 1, 1,
              3,-2, 1,
              2, 1,-1),nrow=3,ncol=3,byrow=T)
d <- c(6,2,1)
solve(A) %*% d
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
```

5. Linear models formulated with linear algebra

Let's consider the previous example of two mice populations in Chapter 3. We have already seen how it connects nicely with t-tests. We will no go through how it can be formulated as using linear regression using matrix formulation aka linear algebra.

Let $x_i = \mathbf{1}(\texttt{Diet = hf})$ which means that $x_i = 1$ if diet is $\texttt{hf}$, and $x_i = 0$ if diet is $\texttt{chow}$.

Lets define notation of matrices and vectors

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots \\ 1 & x_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

So we can write linear regression using the notation above as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and we can use matrix notation, so the above can be written as

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

The $\mathbf{X}$ is what we call a design matrix.

4. Linear hypothesis testing with L-matrices

5. Contrast matrices

6. Confounding