Module 3. Introduction to Statistical Models

BARÇA INNOVATION HUB
Universitas

Module 3. Introduction to Statistical Models

In previous modules we have introduced the types of statistical analyses that there are and that will help us to interpret the data. These analyses may serve different functions, from the description of the data using different measurements, to more complex analyses, such as prescriptive analyses.

Now, we will review the process to follow in data analysis and the steps to follow to benefit from the use of the data collected which are used in decision-making.

- Observation: this has to do with the peculiarities of each of the environments, knowledge of the sport and area of experience of the professional.
- Generation of questions or hypotheses: that is, what we want to fix in order to have a positive impact on our team or on the sporting/academic structure.

- Data collection: the type of design can also be included (it could be experimental if the context allows for it, which is usually the case in the academic or observational field). Choice of tools, data quality and storage.

- Data processing: it has to do with defining the appropriate format for subsequent analysis, deciding how to treat empty data, and add variables or calculations.

- Data analysis: choosing the appropriate tool to address the initial hypothesis.

- Communication/application of results: it involves visualization, extraction of results and decision-making.

This process will be carried out over and over at the end of the last step, in case the new observation is not satisfactory.

The observation and hypothesis steps are not part of this course, but they are key steps within the entire process, especially in data analysis. The same applies to data collection, for which we highlight the importance of collecting data with a purpose, since the success and efficiency of a project depends on the plan for each of the steps.

If necessary you can look at previous modules where we have described the data processing procedure in detail as well as the introduction to visualization, where the principles and bases to communicate results using this tool are highlighted.

Therefore, there is still one fundamental step in the process to be developed: the analysis of the data itself. The choice of the tool that will allow us to respond to the question posed in a precise way. In a previous module we highlighted the types of statistical analyses and the type of functions that each of them have, but it is necessary to delve deeper into these to know the advantages and drawback of each one. In this way, we will be able not only to decide the type of analysis to perform, but also the tool that will be most suitable for it.

When referring to tools we mean statistical tests, some of which we will see below. The type of data we have, the design of the research study and the answer we seek will point to the most appropriate test for analysis.

To illustrate this with an example, if we want to know if our team has improved in jump height during a period of the season, we should choose a diagnostic analysis and, depending on the design of our study, we should choose one tool (or statistical test) or another. However, if the question we want to answer is how the training workload affects the post-session jump values, we will go for a predictive analysis, since we want to use this information to be able

to know the way our players respond or adapt to training stimuli. The tools, again, will also vary depending on the type of data collected and the proposed design.

Particularities of the Sport Scientist' field

Study designs may be divided into two large groups.

Experimental studies: those studies in which we can control the study variables. Modifying and controlling these variables will allow for greater confidence in the results we get. Some of the characteristics of a good experimental study could be splitting the group into a control group and an experimental group or splitting it into random sub-groups.

Observational studies: in most environments of a Sport Scientist, the requirements for experimental designs are not met because of multiple limitations. The schedule does not allow for periods to isolate certain variables that may influence the factor we want to analyse; it is also not common to split the group into two different sub-groups; and as a starting point, we are unable to modify the population or group with which we work. Therefore, we usually collect data on the different variables as they occur.

These factors will affect what is called the internal and external validity of the study conducted. Simply put, internal validity refers to

the guarantee that the relationships between the study variables are causal. The lower the internal validity, the more caution we should have when interpreting results. External validity has an impact on the influence of the results in groups other than the one in our study; therefore, we need to be aware that what may successfully be applied to our team, may not be useful in another group with different characteristics.

In this module, we will focus on the second type of studies, as they are much more common in our environment. Taking into account the above descriptions, we can state that we mostly have observational data.

What is a statistical model?

When we want to answer questions about the sport we work in, the physical performance of our players, or the influence of certain factors on injuries, we are essentially trying to better explain what we observe - the way in which some players differ from others, their similarities, the relationships between physical parameters, etc. We do this in order to predict similar behaviours, anticipate those predictions and take action to achieve the objectives.

Which positions on the team have similar conditional demands? Should they train together or separately?

Which tasks have the greatest influence on the player's cardiac response?

A statistical model is a tool that enables us to understand these relationships. Concepts related to statistics (such as statistical models) seem to be far removed from the daily practical application in the field of the Sport Scientist and may seem to be reserved only to the academic field. We have to move away from randomly determining changes or relationships and rather, use statistical models, which allow us to rigorously use the data we have and make decisions with greater certainty.

A model is essentially a mathematical formula that describes the relationships between a response variable (also called objective/result/dependent variable, or output) and one or more variables that affect that response variable (also called predictive/independent variables, or input).

A model does not intend to explain the complexity of these relationships since it would be impossible to do so in an field such as sports performance, but it aims to simplify the relationships - it is very likely that we do not even know some of the variables that affect the responses we want to know. It describes the essential parameters of the relationships so that estimates can be made in the future (Grolemund; Wickham, 2017).

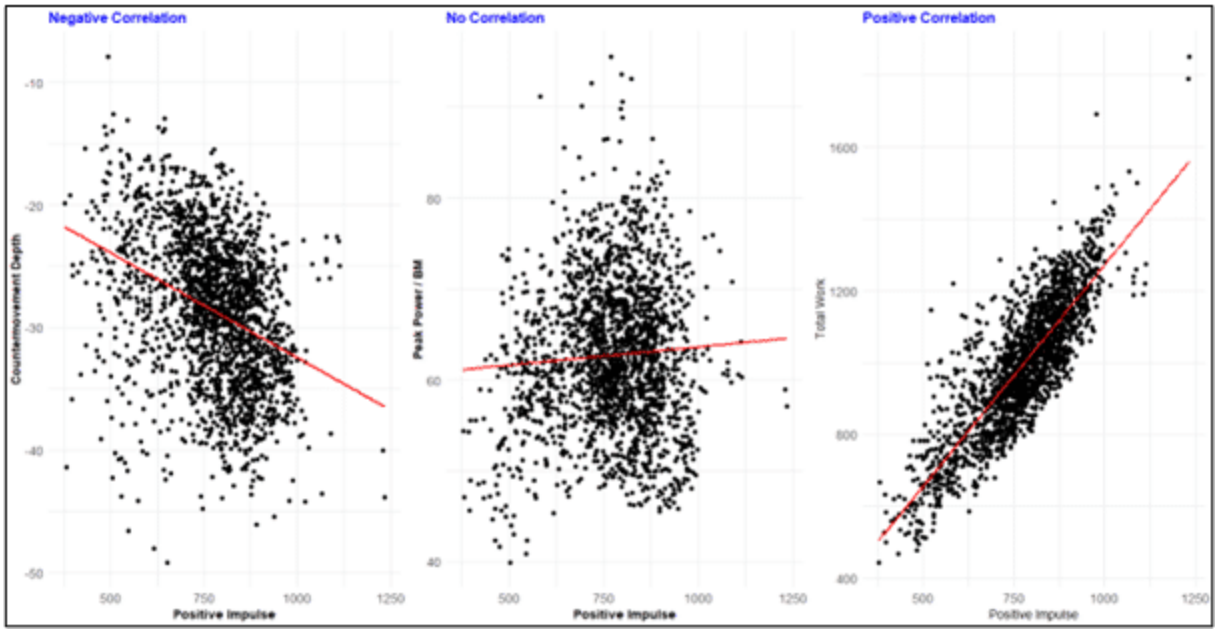
As we have seen in the first module, RStudio is essentially statistical software - it is optimized to meet the needs related to statistical models and has multiple functions to process data in this way.

The linear model

If we believe that two variables are correlated, it means that there is a relationship between them. Therefore, we can use the independent variables to predict the dependent variable.

In the following chart (adapted from Clark, 2020), we can see different types of correlations. It shows the cases in which the variables move together on both axes (rises on the "x" axis are accompanied by rises on the "y" axis in the case of positive correlation).

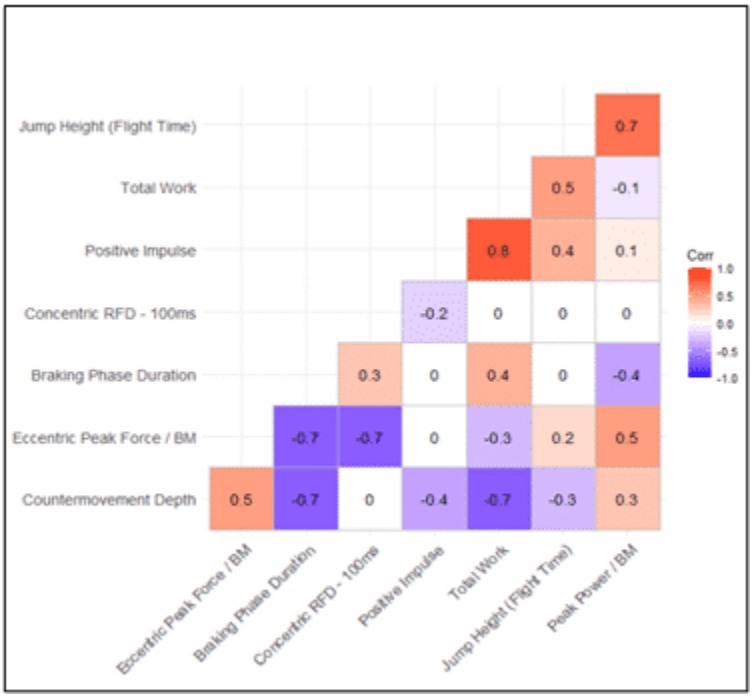
Figure 1: Correlations



Source: author's production based on Clark, 2020.

Correlations are expressed by a numerical value called the correlation coefficient. This number, with a value between -1 and 1, signals the magnitude of the relationship at values closer to 1. In the chart below, there are correlations between different pairs of GPS variables with the numerical value and the corresponding positive or negative relationship also indicated by the colour.

Figure 2: Correlations between different pairs of GPS variables



Source: prepared by the author

These simple correlations suggest a linear relationship, that is, represented by a straight line. This correlation only indicates the direction in which the values move and the magnitude of the correlation, that is, to what extent a change in the value of the x-axis is related to a value on the y-axis. However, this value alone is not of much use if we want to use this relationship to make predictions on new values that we may get. To do this, we need to use a linear model instead.

The linear model is one of the simplest models used to establish relationships between variables. The objective of this model is to find the relationship that indicates the impact of the independent or predictive variables on the response variable and to provide a formula

to which we can add other independent variables and calculate the estimated response variable.

The mathematical formula that represents it is as follows:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

In which:

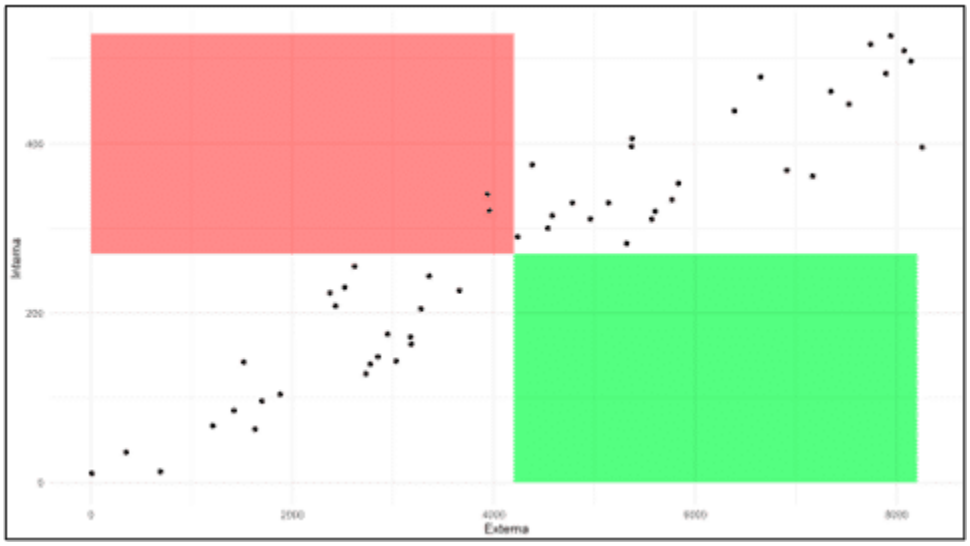
- yy is the response variable;
- x_1, x_2x_1, x_2... they are the independent variables;
- b_0b_0 is the base value;
- b_1, b_2b_1, b_2... are the coefficients or weight of each of the independent variables.

Model creation

We will exemplify the aspects mentioned above. To control of the sports workload, it is of great importance to control both the external load and the internal load of our athletes (Gabbett et al., 2017). Simply put, the external load shows what or how much the athlete is doing (distance, shots, repetitions, etc. depending on the sport we are analysing) and the internal load shows what the stress of that external load is (heart rate, markers of muscle damage, etc.).

The chart below shows different subjects (points) and their two load values, internal on the "y" axis and external on the "x" axis. In this case, we analysed distance in meters as a measure of external load and TRIMPS (Edward's Training Impulse) as a measure of internal load. TRIPS have been calculated by multiplying, in minutes, the duration in each of the heart rate zones by an intensity factor (Tometz et al., 2022).

Figure 3: Different points and values of internal and external load



Source: prepared by the author

The two squares in red and green represent the generalized concept of adaptation or fatigue during training - we could consider that athletes with a high external load and a low internal load (green square) are adapted to training; on the other hand, athletes with a

low external load, but high internal load (red square) may not be adapted or they could be experiencing some fatigue, since there is no efficiency in the cardiac response with the same external stimulus. Although these assumptions may be valid and accurate, we should notice the trend of both variables and know the relationship between them. In this way, we will be able to be more precise in verifying changes and athletes' status.

In addition, using a statistical model may be useful if we want to know the load necessary to induce a certain cardiac response in our athletes.

Following the linear model described above, we will establish that relationship. We want to know the internal load value by knowing the external load value.

Here's the formula above:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

Applied to our example, it would be:

$$Internal\ load = b_0 + b_1 External\ load$$

In RStudio, we will set it up as follows:

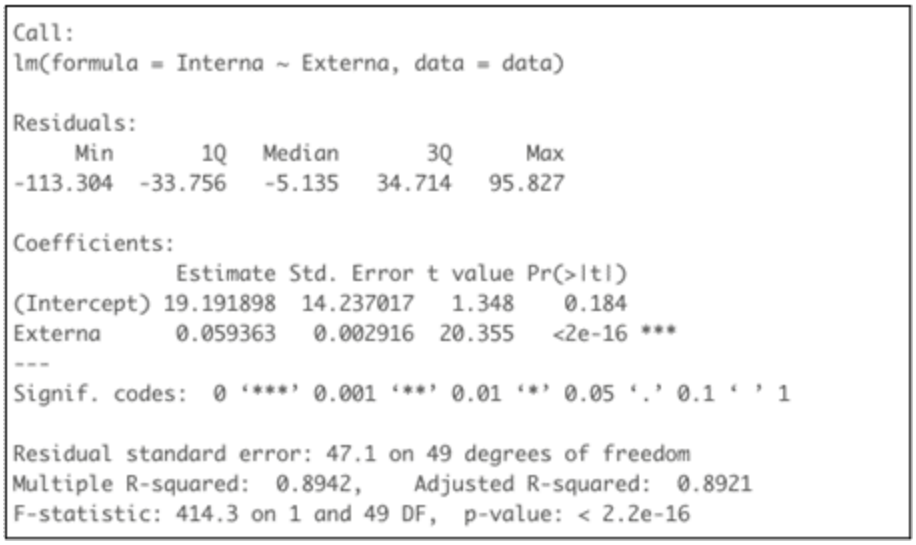
```
model <- lm(Internal~External,data=data).
```

- model: it is the object that collects all the information of the statistical model
- lm(): it is the function that applies the linear model
- Internal and External are the names of the columns in our data source and data is the name of the table
- The variable to the left of the symbol "~" will be the independent variable (what we want to predict), while the variable to the right is the dependent variable (the information we have).

The model in RStudio seems very simple, but it is all we need. From now on, what we must take into account are the results of this model, which will help us assess whether it is useful and reliable to make predictions.

Using the summary() function we see the results presented by RStudio as shown in the following chart.

Figure 4: RStudio results



```
Call:
lm(formula = Interna ~ Externa, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-113.304  -33.756   -5.135   34.714   95.827

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.191898  14.237017   1.348    0.184
Externa      0.059363   0.002916  20.355   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.1 on 49 degrees of freedom
Multiple R-squared:  0.8942,    Adjusted R-squared:  0.8921
F-statistic: 414.3 on 1 and 49 DF,  p-value: < 2.2e-16
```

Source: prepared by the author.

From all the information we have, we will look, for now, at the "Coefficients" section. These are going to complete the part of the formula that we are missing. **We see in the "Estimate" column that there are values for the "(Intercept)", it is the b_0 b_0 value of our formula, and for "External"**. In this way, the formula would be as follows.

Internal load = 19.19 + 0.059xExternal load

In this way, we can already make some predictions. The coefficient in this linear model indicates that for every 1 unit increase in external load, there will be an added increase of 0.059 in internal load.

Therefore, if we want to estimate the internal load of a session in which the athlete has covered 5 km (5000 m), the formula will be as follows.

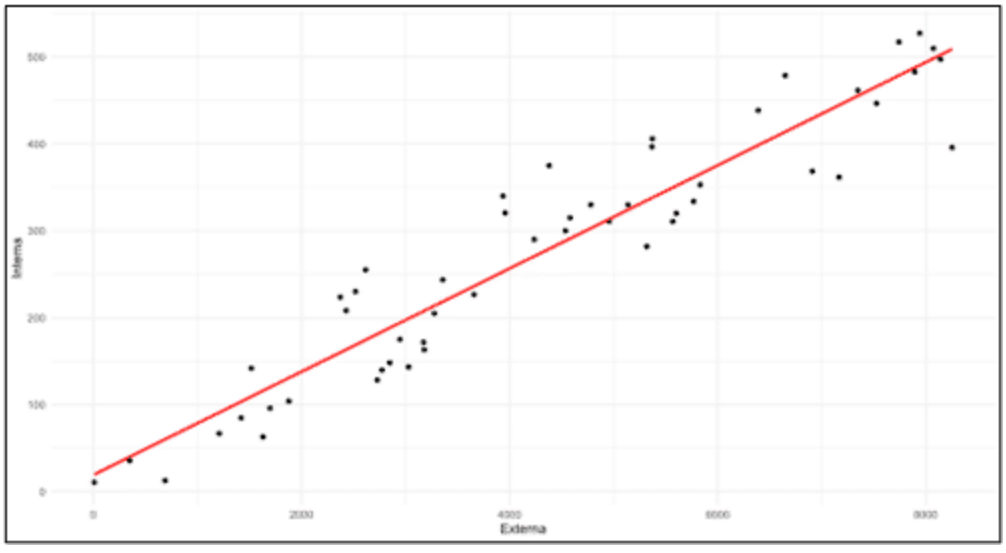
Internal load = 19.19 + 0.059x5000

Internal load = 314.19

In the video material we will see how to make predictions on new data or higher volumes, but this is the basis of the reasoning behind all linear models.

Visually, when there is a dependent variable it can be represented with a line, as in the following chart.

Figure 5: Dependent variable representation



Source: prepared by the author

This line represents the estimated internal load values for each internal load value. As we can see, the line is supposed to represent in the best possible way the way in which the data is presented. The models are not perfect, which is why the red line does not follow all the points but it is intended to make the differences between the real value and the estimated value as small as possible. This concept is called error.

Model assessment

We have just mentioned that all models have a certain error, since we always seek to understand the relationships of complex elements in our environment. When developing models we should try to keep this error as low as possible.

The error is represented by the "residuals", which are the differences between the current values and those estimated by the model. The following chart shows the differences.

Figure 6: Differences between current values and those estimated by the model

Source: prepared by the author

In the results we have seen before, the error is represented by the value "Residual standard error", which shows the usual distance between the points, in this case 47.1, which is a very low value when considering the internal load units.

Another element to assess our model is the R-Squared value. This value, always between 0 and 1, will indicate what proportion of the variable we want to predict is explained by the independent variable. In this case, the value is close to 1, which indicates good results.

This model opens new possibilities for the initial analysis that we had proposed. We know the formula for estimating new values, we know the error of our model and, if we have done a correct exploration of our data, we will also know the variability of our measurements. With these aspects in mind, we will be in a better position to assess our

athletes' responses. Values further away from the estimate will indicate different responses to training and we may leave aside the square analysis that we have shown at the beginning of this section.

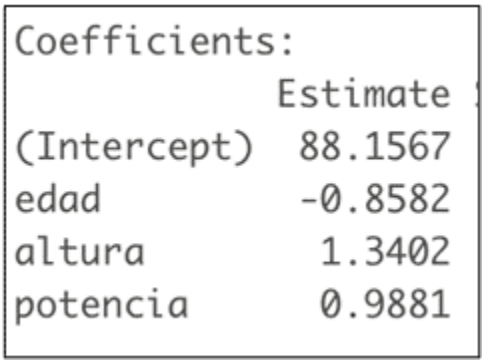
Model complexity

The linear model and the regression that we have shown in the example belong to the simplest model - we need to understand the data we are dealing with, carry out the corresponding exploration and elaborate the question we want to respond and then choose the type of model that best fits those data to get the desired results.

Previously, we have shown a model with a single dependent variable, in which the results were quite good. However, in linear models we usually want to use more than one variable to estimate the response variable. Following the same formula presented above, we will add more variables with their coefficients corresponding to each of them. In the same way that we have interpreted the previous model to make estimates, we would do so with the following ones.

For example, we could have a model where we wanted to estimate the speed of a tennis shot. The variables we could use for this example are the player's age, height, and a value called "power," which could represent values got by this athlete in a strength test.

Figure 7: Example



Source: prepared by the author

The formula in this case would be as follows.

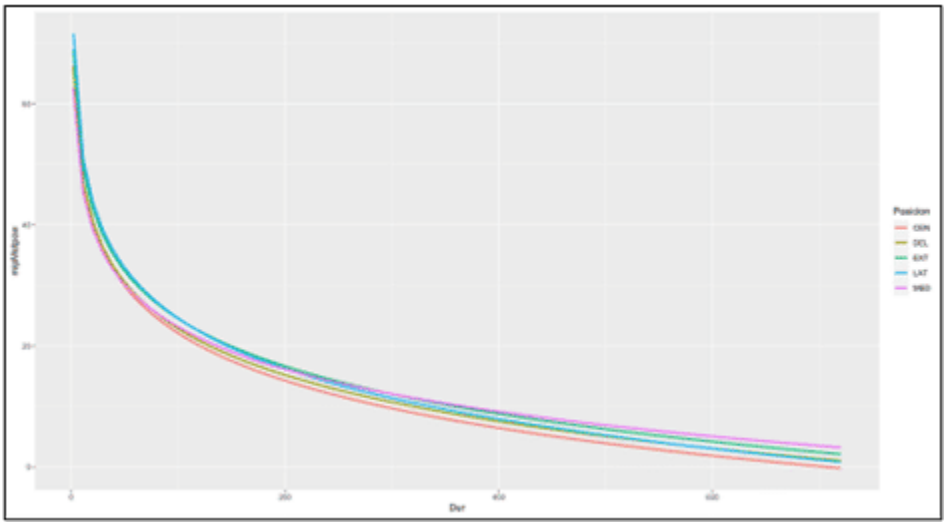
y= 88.15 -0.85xage +1.34xheight +0.98xpower

This is just an example with non-real values and variables, but it has been used to observe the negative coefficient in the model. This represents how age has a negative impact on the values we will estimate. In other words, the model would indicate that an older athlete, with the same height and power values as a younger athlete, will have a lower service speed.

We could also consider linear models in which the line that relates both variables is not straight.

An example would be periods of maximum intensity. These periods, already used in some other examples of the course, show those phases in which an athlete expresses intensity values (in this case reflected by the metabolic power variable) higher than the average values in the match. We can also simplify them and refer to speeds. The horizontal axis represents the duration. As we can see, the shorter the duration of the athlete's effort (in this case, of a professional football team) the greater the intensities they are able to reach: as the duration increases those intensities are lower. The same concept applies to speed: we are unable to keep the same speed in a 6-second effort as in a 2-minute effort.

Figure 8: Example of periods of maximum intensity



Source: prepared by the author

As described by Delaney et al. (2018), this relationship is expressed by a linear model in which we need to perform certain transformations to the variables to get the model formula and make predictions using the corresponding formula.

There is also a chance that the variable we want to describe is not numerical. For example, we want to build a model to predict whether a tennis serve will be a straight serve or not. In this case, the prediction will be binary (there are two possibilities, yes or no). To do this, we want to use several independent variables (speed, location, effect, etc.). Iit is clear that the independent variables also have particularities: not all of them are numerical variables but there are categorical variables as well - the effect has limited categories, for example (topspin, flat, cut). In many cases, the distribution of the variables is not normal, nor are those of the response variable. For all these cases, it is necessary to use generalized linear models. These models enable us to work with the variety of data we have described. Even if the function used in RStudio is different (glm()), the construction of the model will be practically the same. Below there is another example.

We will try to predict whether a serve has been direct or not (ace) using the variables:

- speed of the serve (speed)

- effect (effect). The effect has 3 categories:
    - flat (flat serve);
    - slice (slice serve);
    - spin (topspin serve).

We elaborate the model with the formula shown below, in addition to the function that specifies that it is a generalized linear model (glm), and we add that it is a binary response variable (family="binomial"). The results we would get are as follows.

Figure 9: Results

```
Call:
glm(formula = ace ~ speed + effect, family = "binomial", data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0140  -0.8891  -0.3794   0.9222   2.3932

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.51644    2.20016  -3.871 0.000108 ***
speed        0.05007    0.01221   4.101 4.12e-05 ***
effectslice -0.67241    0.43487  -1.546 0.122045
effectspin  -1.93588    0.53402  -3.625 0.000289 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 193.82  on 139  degrees of freedom
Residual deviance: 158.40  on 136  degrees of freedom
AIC: 166.4
```

Source: prepared by the author.

In this case, we can see how the model assigns coefficients or weights to each of the categories. By default, the model omits one of the categories, which serves as a reference, but the reasoning is the same. Also, as it is a binary model, the results will be in the format of probability that one condition or another is met and we have to choose what the threshold is to determine if it has been a direct serve or not. These models are usually assessed through their accuracy, that is, the success they have had in determining positive and negative cases.

There are functions in RStudio that allow you to make predictions, use the formulas of the models and efficiently assess their performance

without going into detail, but it is important to have a clear idea of each of their objectives.

Knowing the particularities of linear models is the basis for the construction of more complex models.

CONTINUE

References

Clark, M. (2020). *Model estimation by example. Demonstrations with R.* https://m-clark.github.io/models-by-example/

Delaney, J. A.; Thornton, H. R.; Rowell, A. E.; Dascombe, B. J.; Aughey, R. J.; Duthie, G. M. (2018). Modelling the decrement in running intensity within professional soccer players. *Science and Medicine in Football*, 2(2), 86–92. https://doi.org/10.1080/24733938.2017.1383623

Gabbett, T. J.; Nassis, G. P.; Oetter, E.; Pretorius, J.; Johnston, N.; Medina, D.; Rodas, G.; Myslinski, T.; Howells, D.; Beard, A.; Ryan, A. (2017). The athlete monitoring cycle: a practical guide to interpreting and applying training monitoring data. *British Journal of Sports Medicine*, 51(20), 1451-1452. https://doi.org/10.1136/bjsports-2016-097298

Grolemund, G.; Wickham, H. (2017). R for Data Science. O'Reilly Media

Tometz, M. J.; Jevas, S. A.; Esposito, P. M.; Annaccone, A. R. (2022). Validation of Internal and External Load Metrics in NCAA D1 Women's Beach Volleyball. *Journal of Strength and Conditioning Research*, 36(8), 2223-2229. https://doi.org/10.1519/JSC.0000000000003963

CONTINUE